BIOINFORMATICS ANALYSES OF NON-CODING GENOMIC ELEMENTS

BY

KAI ZHAO

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Animal Sciences
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Doctoral Committee:

Associate Professor Alfred L. Roca, Chair
Assistant Professor Felipe C. Cardoso
Associate Professor Juan J. Loor
Associate Professor Saurabh Sinha

ABSTRACT

Mammalian genomes consist primarily of non-coding sequences (Kellis *et al.* 2014). Originally castigated as "junk DNA", many non-coding regions have now been characterized as having functional roles, or have been determined to be the causal agent for diseases. Additionally, sequences that are non-functional can be used as neutral markers for population genetics. Determining the role of non-coding sequences or finding sequences usable as neutral markers is computationally and biologically non-trivial. However, recent advances in molecular biology, in particular the reduced cost of next-generation sequencing (NGS), have enabled new experiments that involve these sequences. I will discuss studies using bioinformatics that leveraged these advances to characterize three types of non-coding sequences: endogenous retroviruses, microsatellite markers and transcription factor binding sites. I conducted the bioinformatics design, coding and analyses, working with collaborators who verified findings in the laboratory.

The only retrovirus known to be currently transitioning from exogenous to endogenous form is the koala retrovirus (KoRV), making koalas (*Phascolarctos cinereus*) ideal for examining the early stages of retroviral endogenization. In the first study, I developed a bioinformatics routine to identify distinct retroviral integrants from NGS reads of KoRV retrovirus flanks isolated using koala genomic DNA. In the second study, I developed computationally efficient, user-friendly software that would identify polymorphic microsatellite loci using NGS reads, then design oligonucleotide primers appropriate for amplifying those loci. We developed this software to enable studies to improve understanding of population structure, estimate population size and estimate genetic diversity in genetically depauperate wildlife species. In the third study, I

developed a bioinformatics pipeline to characterize gene expression changes during development in the fetal limb tissue of several mammalian species, to better understand the mechanistic differences across evolutionary lineages. We compared development in four species of mammals. The house mouse was used since it is a well-characterized model organism with five digits. The domestic pig was used since it is a well-studied agricultural animal and a model for digit reduction. A species of bat was used since bats undergo wing development. Finally, a species of opossum was used as an outgroup to the three eutherian species.

To my 爷爷奶奶姥爷姥姥 (grandparents)—

ACKNOWLEDGEMENTS

I would like to thank my advisor, Prof. Alfred Roca, for your continued and tireless support. I cannot thank you enough for devoting your time, energy and funding to cultivate me from a fledging undergraduate student. It could not have been easy and I am eternally indebted for your role in developing me as person, teacher and scientist.

Many thanks to Prof. Saurabh Sinha for giving me indispensible lessons in bioinformatics. I also thank Prof. Jian Ma, Prof. Juan Loor, Prof. Phil Cardoso and Prof. Anna Kukekova for graciously serving on my committee and giving me valuable feedback for my research. In addition, thanks to Prof. Karen Sears and Dr. Jennifer Maier for the opportunity to collaborate on the limb development project.

My colleagues in the Roca lab, for your crucial roles in the projects I've worked on and for making the lab feel like my second home— Yasuko Ishida, Jess & Adam Brandt, Tolu Stowe, Christina Ruiz and Alida de Flamingh.

I truly enjoyed my many experiences teaching here at the University of Illinois— Thanks to Prof. Tom Gambill, Prof. Viraj Kumar, Prof. Steven Lauterberg, Prof. Craig Zilles, Prof. Wade Fagen and Prof. Amy Fischer for giving me these opportunities and having the faith in me to teach your students. I would also like to thank Holly Bagwell and the rest of the academic staff in the Computer Science Department for their untiring support.

The many years I spent here in Champaign-Urbana would have certainly felt longer without my friends: Majid Kazemian, Ryan Cunningham, Han-Wen Yeh, Dan Lior, Jason Cho, Thyago Duque, Kyle Tsai, Halie Rando, Diana Byrne, Boris Sadkhin, Laura Sloofman, Matt Foreman, Melissa Chua, Mary Hudson, Meiling Liu, Maria Zyskind and Erik Wang. We shared countless memories traveling, teaching, cooking and just hanging out together. And special thanks to Guy Tal, my roommate of many years, with whom I taught together for five semesters and my partner in numerous projects, who above all, taught me *joie de vivre*.

Yoo Min—Your loving and unwavering encouragement has been pivotal in giving me the energy and the perseverance to finish this thesis and starting my career.

Mom and Dad, thank you for serving as the role models for me and helping me to put education first. On the cusp of getting this doctorate, I finally can comprehend its importance to me and my future.

Portions of the work presented in this dissertation were supported by the International Rhino Foundation, the US Fish and Wildlife Service Rhinoceros and Tiger Conservation Fund, and the National Institute of General Medical Sciences (NIGMS) grant number R01GM092706.

TABLE OF CONTENTS

CHAPTER 1: IDENTIFICATION OF KOALA RETROVIRUS FLANKING REGIONS

Introduction

Endogenous retroviruses (ERVs) are pervasive in the genomes of all vertebrate lineages, and comprise approximately 8% of the human genome (Bromham 2002; Lander *et al.* 2001). ERVs originate from exogenous retroviruses that integrated into the ancestral host germ line and were subsequently passed from parent to offspring through Mendelian inheritance (Bromham 2002; Coffin 2004; Stoye 2012). Most ERVs are neutral or deleterious to the host; they decay into non-functional sequences over time through mutation. However, ERVs have been shown to recombine with other endogenous or exogenous viruses, protect the host against similar exogenous viruses, retain the ability to produce viral protein, or even become co-opted into a functional role for the host (Bromham 2002; Coffin 2004; Stoye 2012). *Syncytin*, a gene that plays a vital role in normal human placentation, is derived from a retroviral envelope gene that integrated into the germ line following an ancient infection (Mi *et al.* 2000). Conversely, de-repression of a human ERV was found to facilitate Hodgkin's lymphoma (Lamprecht *et al.* 2010). Phylogenetic studies of endogenous retroviruses reveal that retroviruses have frequently jumped from one species to another and integrated into the germ lines of their hosts (Denner 2007; Fiebig *et al.* 2006; Hayward *et al.* 2013).

Since most ERVs integrated into host genomes millions of years ago, it is difficult to characterize the mechanisms involved in a germline invasion (Coffin 2004; Johnson & Coffin 1999; Stoye 2006). The koala retrovirus (KoRV) has recently been identified as an extraordinary instance of a virus in the midst of endogenization. Koala populations in

northern Australia exhibit 100% prevalence of KoRV, carrying an average of 165 copies per cell, while in southern Australian populations many koalas are completely free of the virus (Simmons *et al.* 2012; Tarlinton *et al.* 2006). This suggests that KoRV initially affected koalas in northern Australia and is currently spreading to southern populations (Tarlinton *et al.* 2008; Tarlinton *et al.* 2006). There also appear to be KoRV variants with more limited distributions that may be of more recent origin (Shimode *et al.* 2014; Shojima *et al.* 2013a; Shojima *et al.* 2013b; Xu *et al.* 2013).

KoRV currently exists as an endogenous retrovirus, but is also thought to be transmitted horizontally (Shimode *et al.* 2014; Shojima *et al.* 2013a; Shojima *et al.* 2013b; Stoye 2006; Tarlinton *et al.* 2008; Xu *et al.* 2013). Previous studies have suggested that KoRV exists in both an endogenous and exogenous state (Stoye 2006; Tarlinton *et al.* 2006). One issue in interpreting past studies of KoRV has been that the proviruses of KoRV that were detected could have been endogenous or exogenous. In this study, we isolated KoRV flanking sites in the koala genome using a modified genome-walking approach (Reddy *et al.* 2008) and we developed a bioinformatics technique to reconstruct integration sites of KoRV. We used the reconstructed integration sites to determine whether KoRV proviruses in the genome are endogenous, by establishing Mendelian inheritance using a sire–dam–progeny triad of northern Australian (Queensland) koalas kept in North American zoos. A provirus found at a particular locus in the progeny would be established as endogenous if it was also found in either parent at the same locus, as two ERVs independently integrating at the same locus in two individuals would be an extremely rare event (Johnson & Coffin 1999).

Methods

KOALA SAMPLES

Ethical approval for this study was granted by the University of Illinois Institutional Animal Care and Use Committee, approved protocol number 12040. Blood samples from northern Australian koalas were obtained during regular physical examinations by trained staff at the Columbus Zoo and the San Diego Zoo, USA. The American Zoo Association's Species Survival Plan manages northern (Queensland) and southern koalas separately. Three northern Australian koalas comprised a parent–progeny triad (progeny: Pci-SN404, sire: Pci-SN248, and dam: Pci-SN345). The pedigree of these individuals was available in the North American Studbook for koalas. Inbreeding was known to be limited in their pedigree. The parents shared only a single distant ancestor (great grandparent to the sire and great–great grandparent to the dam) and thus had a low estimated relatedness ($r \cong 0.008$). In addition to the triad, other zoo samples that were kin to the triad included Pci-SN374, the daughter of Pci-SN248 and Pci-SN345; and two patrilineal siblings of Pci-SN345: Pci-SN339 and Pci-SN356. For zoo koalas, genomic DNA was extracted from buffy coat using the QIAamp DNA Blood Mini Kit (Qiagen, Valencia, CA). The southern Australian koala DNA samples were provided by the National Cancer Institute (NCI), USA, used the NCI sample numbers, and had been collected from free-ranging wild koalas in Australia. Pci-157 was from the Stony Rises of Victoria, Pci-106 was from the Brisbane Ranges of Victoria, and Pci-187 was from Kangaroo Island of South Australia (Table 1.1). The blood samples had been collected under permit no. 87-150 issued by the Department of Conservation, Forests and Land,

Victoria (Taylor *et al.* 1991). The DNA had been extracted using a phenol–chloroform method at NCI.

Screening Koala Samples for the Presence of KoRV

As some koala individuals and populations are largely free of KoRV, the DNA samples used in this study were screened to determine that they were KoRV positive by using PCR primers that were previously published (Tarlinton *et al.* 2006) or newly designed based on conserved regions of the LTRs (3'-LTR-F2: AGTTGTGTTCGCGTTGATCC, KoRV-3LTR_F2R: TACCTCCCGTCGGTGGTT). The primer 3'-LTR-F2 was also used to isolate KoRV flanking regions (the next section has details). The PCR setup is described below, whereas the algorithm used was as previously described (Ishida *et al.* 2011).

Isolation and Sequencing of Koala Genomic Regions Flanking KoRV Proviruses

To identify host genomic regions flanking KoRV proviral integration sites, the genome-walking method established by Reddy et al. (2008) was implemented, although modified to use next-generation sequencing as illustrated in Figure 1.1. The REPLI-g Mini Kit (Qiagen) was used. Approximately 100 ng of each koala genomic DNA was denatured, following the REPLI-g kit protocols. Four different walker-adapter primers were then attached to each sample of denatured DNA (Reddy *et al.* 2008) (Table 1.5), using a mix that consisted of 10 units of Phi29 DNA polymerase, 1× Phi29 DNA polymerase reaction buffer, 200 µM dNTPs, and 20 µM of each walker-adapter primer.

The mixture was incubated at 30°C for 90 min to initiate multiple primer extension events, then incubated at 65°C for 10 min to inactivate the polymerase. The QIAquick PCR Purification Kit (Qiagen) was used to remove unincorporated walker-adapter primers, following the manufacturer's protocol. The purified DNA fragments with walker-adapter primers were eluted using 40 µl of TLE buffer.

The eluted DNA was used as template for PCR procedures involved in the genome-walking method (Figure 1.1) (Reddy *et al.* 2008). Each PCR relied on one walker primer and one KoRV-specific primer (Table 1.5). The KoRV-specific primers were designed using Primer3 (Rozen & Skaletsky 2000) and designed to target the 5'-end or 3'-end of the LTR based on regions conserved among published KoRV sequences available at the time: GenBank accession numbers AF151794 (Hanger *et al.* 2000), DQ683164, DQ683166, DQ683167, and DQ683168 (Tarlinton *et al.* 2006). A primary PCR was conducted as previously described (Tarlinton *et al.* 2006). In the subsequent nested PCR, the primary PCR product was used as template, and amplified using a pair of HPLC-purified primers (Integrated DNA Technologies, Coralville, IA). Primers were prepared by following the manufacturer's protocol for the Roche Genome Sequencer System (Roche Applied Science, Penzberg, Germany). One primer consisted of three concatenated segments: A GS FLX Titanium adapter "Primer A" segment (CCATCTCATCCCTGCGTGTCTCCGACTCAG), a MID, and a KoRV-specific primer (Table 1.5). The MID used was the same across the four different amplicons of walker-adapter but distinctive for each koala individual, and for each run (5' or 3'). The second primer consisted of two concatenated segments: The GS FLX Titanium adapter "Primer B" (CCTATCCCCTGTGTGCCTTGGCAGTCTCAG) and a second walker primer

previously described (Reddy *et al.* 2008) (Table 1.5). PCR was conducted using the FastStart High Fidelity PCR System (Roche Applied Science) and the PCR components and algorithm conformed to the manufacturer's protocol. The resulting PCR amplicons were purified using AMPure XP beads (Beckman Coulter, CA) with a magnetic particle concentrator. The concentrations of the purified nested PCR amplicons were estimated using a Qubit 2.0 Fluorometer (Life Technologies Corp.) and amplicon sizes, quality, and quantity were measured using an Agilent 2100 Bioanalyzer at the Functional Genomics Unit, Biotechnology Center (Biotech Center) at the University of Illinois at Urbana-Champaign (UIUC). Amplicon concentrations were adjusted so that equal amounts would be pooled. The pooled sample was eluted on an agarose gel and separated into two size classes, one approximately 200–400 bp and the other approximately 400–1,000 bp at the High-Throughput Sequencing and Genotyping Unit, Biotech Center at UIUC. Each size class was run separately on 1/16th of a PicoTiterPlate (PTP) (1/8 PTP total) on the Roche 454 GS FLX+ platform at the UIUC High-Throughput Sequencing and Genotyping Unit.

BIOINFORMATICS PROCESSING OF NEXT-GENERATION SEQUENCES

Reads generated by the Roche 454 GS FLX platform were converted into FASTQ format using the Galaxy bioformatics platform(Giardine *et al.* 2005). The experimentally ligated MID formed part of the sequence read and indicated which koala the sequence originated in, and whether the 5'-end of the LTRs or the 3'-end of the LTRs was the target. As 5'- and 3'-LTRs have nearly identical sequences, about half of the PCR amplicons and subsequent sequencing reads would be expected to identify sequence

within the KoRV provirus rather than sequences in the host flanks. To remove reads matching the KoRV provirus, we used Bowtie2 (Langmead & Salzberg 2012), using the "very sensitive local alignment" preset, to attempt to map all reads to published KoRV sequence AF151794 (Hanger *et al.* 2000). Only reads that did not map to KoRV genes were further considered.

To identify the boundary between the KoRV LTR and the koala flanking genomic sequence, the flanks were mapped onto the published KoRV LTR sequence using Bowtie2 (Langmead & Salzberg 2012), using the "very sensitive local alignment" preset. A number of steps were taken to find the matching 5'- and 3'-flanks at a single proviral locus. First, the flank sequences were trimmed to include approximately 10 bp of the end of the LTR and 10 bp of the koala genomic regions. Each flank sequence was aligned to the Meug_1.1 assembly of the genome of the tammar wallaby (Renfree *et al.* 2011) using BLASTN (Altschul *et al.* 1990) using parameters for short local alignment. Flanks that aligned to more than three scaffolds were removed to reduce the possibility that multiple unique flanks of KoRV might be misidentified as one insertion. We wrote a routine using BioPython (Cock *et al.* 2009) to filter the BLAST results for pairs of 5'- and 3'-koala genomic flanks. Pairs of 5' and 3'-flanks that aligned to the same wallaby scaffold, with a -10 bp to 10 bp overlap, were assumed to be from the same host integration site. The quartet of sequences (5' of the proviral integration site, 3' of the proviral integration site, the matching wallaby segment) and the published KoRV sequence (Hanger *et al.* 2000) was then realigned and visually inspected in the software Sequencher 5.1 (Gene Codes Corp., MI). Since we expected each LTR pair to originate from a single integrant, the first 4-6 bp immediately flanking the 5' and 3' of the retroviral sequence should be

identical. Surprisingly, we did not find this to be the case in *any* of the ten KoRV LTR pairs. We determined that removing the first 2 bp of the 5' LTR and 1 bp of the 3' LTR of the published retrovirus sequence, allowed a 4 bp target site duplication to be identified in nine of the ten LTR pairs (one LTR pair had a 5 bp target site duplication). Correcting the 2 bp and 1 bp anomalies in the published KoRV sequence was imperative for the remainder of the study—to reveal the size of the KoRV target site duplication and to facilitate correct trimming of the majority of KoRV flanks that did not align to the wallaby genome.

To identify additional matched flanking sequences on either side of a single proviral locus, all flank sequences were queried against low-coverage koala genomic sequences. For this search, Bowtie2 (version 2.1.0) (Langmead & Salzberg 2012) was run on the Galaxy platform (Giardine *et al.* 2005). The koala genomic reads had been generated using DNA from Pci-SN404, sequenced on 1/16th of a PTP of the Roche 454 GS FLX+ platform (Roche Applied Science) run at the High-Throughput Sequencing and Genotyping Unit, UIUC, as has been previously described (Ruiz-Rodriguez *et al.* 2014).

To estimate the number of distinct retroviral integrations from the host flanks sequenced by the Roche 454 GS FLX platform (Table 1.3), the reads were trimmed to only include approximately 50 bp of host genomic flank adjacent to the proviral LTR. We retained only those reads that contained at least 50 bp of host genomic flank and had a base call quality of 99% for every position in the 50 bp. This minimized the possibility of an inflated count due to sequencing errors. For each MID data set iteration, we used the Megablast algorithm in BLASTN (Altschul *et al.* 1990) to cross-align all filtered reads from the same iteration, and grouped together those reads that were at least 80%

similar, with each group of reads counted as a "distinct" flank sequence. These criteria for grouping the number of distinct flank sequences may have somewhat underestimated the total. For each distinct grouping of flank sequences, the consensus 4 bp at the LTR boundary was taken as the target site duplication for the integration site. The number of proviruses for each koala was estimated as the number of distinct flank sequences detected for 5'- and 3'-flanks separately, and present in at least two of the sequence reads (singletons were removed to minimize potential error). Then for each target site duplication, the number of 5'- and 3'-distinct sequences was compared, and the larger of the two for each target site duplication was used in estimating the total number of reads for each individual koala (Table 1.3). The number of sequencing reads per distinct flank is shown in Figure 1.2.

PCR AND SEQUENCING OF FLANKS AND LTRS

PCR primers were designed using the software Primer3 (Rozen & Skaletsky 2000) targeting koala genomic sequences flanking proviral integration sites, or targeting KoRV LTR sequence (Tables 2.6 and 2.7). Primers for identification of enKoRVs in the dam–sire–progeny triad were designed based on flank reads from the Roche 454 GS FLX+ platform for Pci-SN404 (progeny). To minimize potential bias in detecting endogenous over exogenous KoRVs, half of the primer sets were designed based on distinct flanks that were detected in high frequencies among the sequence reads, whereas the rest were designed based on distinct flanks that were detected in low frequencies among the sequence reads (Table 1.4). Only the successful primers are shown in Table 1.7. To minimize the targeting of repetitive regions within the koala genome, flank

primer sequences were queried against low coverage whole-genome sequence of Pci-SN404 from a 1/16th PTP run on the Roche 454 GS FLX+ platform (Ruiz-Rodriguez *et al.* 2014), although none of them was found to be in repetitive regions by using this low coverage sequence. When the same 4-bp target site duplication was identified upstream of a 5'-LTR and downstream of a 3'-LTR, PCR was conducted using a primer that targeted the 5'-flank with one that targeted the 3'-flank, to determine whether the two primers flanked the same locus, using DNA from a koala known not to carry the relevant KoRV(s).

PCR mixes used a final concentration of 0.4 µM of each primer, 1.5 mM $MgCl_2$, 200 µM of each dNTP (Life Technologies Corp., CA), and 0.04 units/µl of AmpliTaq Gold DNA Polymerase (Life Technologies Corp.). The PCR algorithm consisted of an initial denaturation and activation of AmpliTaq Gold at 95 °C for 9:45 min; with cycles of 20-s denaturation at 94 °C, followed by 30-s annealing at 60 °C (first three cycles), decreasing the annealing temperature in 2 °C steps to 58, 56, 54 and 52 °C (five cycles each), or 50 °C (last 22 cycles), followed by 1-min extension at 72 °C; with a final extension of 7 min at 72 °C. An aliquot of each PCR amplicon was examined on an agarose gel with ethidium bromide under UV light. Amplicons were treated with Exonuclease I (USB Corporation, OH) and shrimp alkaline phosphatase (USB Corporation) to remove excess primers and unincorporated dNTPs (Hanke and Wink 1994). Sanger sequencing was performed in both directions using the BigDye Terminator v3.1 Cycle Sequencing Kit (Life Technologies Corp.) with 2.5 µl of purified PCR product and 0.12 µM primer (M13 forward or reverse), as previously described (Ishida et al. 2011), and purified and resolved on an ABI 3730XL capillary sequencer at the High-

Throughput Sequencing and Genotyping Unit, Biotech Center at UIUC. The software Sequencher 4.5 (Gene Codes Corp., MI) was used to examine and edit chromatograms.

For each distinct proviral flank verified by PCR and Sanger sequencing, the 50 bp of host genomic sequence flanking the provirus identified by Sanger sequencing was used as a query against the Roche 454 flank sequencing data set, and the number of matching reads was recorded (Table 1.4), in order to show that proviruses were evenly distributed among flanks with low numbers of reads and flanks with high numbers of reads.

Results and Discussion

We sought to identify KoRV integration sites in six KoRV positive koalas, three from northern and three from southern Australia. The three northern Australian koalas were from zoos (Table 1.1a), comprising a sire-dam-progeny triad (offspring: Pci-SN404, sire: Pci-SN248, and dam: Pci-SN345). The three southern Australian koalas were unrelated, wild-caught and chosen for the diversity of their geographic origins (Table 1.1b). One koala each was from the Stony Rises (Pci-157) and the Brisbane Ranges (Pci-106) of Victoria, and Kangaroo Island (Pci-187) of South Australia.

To identify host genomic DNA flanking the 5' and 3' KoRV LTRs in each koala, a genome-walking method (Reddy *et al.* 2008) was implemented, but modified to use next-generation sequencing, as illustrated in Figure 1.1. The flanks were sequenced using the Roche 454 GS FLX+ platform. A unique multiplex identifier (MID) was used for each flanking sequence of each koala, generating 12 sets of sequences. A total of 136,430 reads were generated across the koalas. The number of reads was high for all attempts on

the triad (Table 1.1), as was the average percentage of reads that contained the koala genomic flanks (31 - 48%). This average was much lower (< 1%) for 3 of the 6 attempts on southern koalas—the 5' attempt for Pci-106, and both flanks for Pci-187. The reason for this reduced success was unclear, although KoRV is less common in southern koalas. Since the target of the study was the sire-dam-progeny triad of northern Australian koalas, for which all attempts were very successful, the genome walking method was not repeated for less successful southern koalas.

The koala genomic sequences flanking KoRV integration sites were queried against genomic scaffolds of the Meug_1.1 assembly of the tammar wallaby (*Macropus eugenii*) genome (Renfree *et al.* 2011). In some cases, 5' and 3' flanking sequences were found to match adjacent regions of the wallaby genome, suggesting that the two flanks would correspond to koala genomic sequence on either side of the integration site of a KoRV locus. Comparison of the 5' and 3' host genomic flanks for a provirus at a single locus permitted identification of the "target site duplication" on either side of the provirus. The "target site duplication" is a region of host DNA that is replicated during integration of a retrovirus, so that the same 4-6 bp sequence appears immediately upstream and downstream of the integrated provirus. We determined that the length of the target site duplication is 4 bp for KoRV.

The number of KoRV integration sites detected in the 3 northern koalas was 74 for the sire, 69 for the dam and 105 for the progeny. Among southern koalas, Pci-157 had a count of 16, while Pci-106 had a count of 10 (with only one flank successful) (Table 1.1a), consistent with lower copy numbers for KoRV previously reported for southern koalas. Given the stringent criteria used in the bioinformatics approach and, the poor

quality of many reads, these numbers likely underestimate the number of distinct flanks. The mismatch in counts between the 5' and 3' flanks for each koala also indicated that the method did not identify all flanks comprehensively.

The target site duplication was used to designate individual proviral loci. For example, if the target site duplication on either side of the KoRV provirus had a sequence of ACGT, the provirus was designated "KoRV-ACGT." Subsequent PCR and sequencing of individual proviral loci (below) confirmed the 4 bp length of the target site duplication. There was one exceptional provirus that had a 5 bp target site duplication, KoRV-AAAAG. The integration site for this provirus included four adenine bases in tandem, suggesting that the longer target site duplication may have resulted from strand slippage of the host DNA (Ballandras-Colas *et al.* 2013; Craigie & Bushman 2012; Levinson & Gutman 1987), although other target site duplications were 4 bp in length despite the presence of homopolymers (e.g., AAAG, AAAT, CCCC).

We identified both the 5' and the 3' koala genomic regions flanking the KoRV integration site for 10 loci (Table 1.2), mostly by homology to the tammar wallaby genomic sequence. The wallaby and koala lineages diverged more than 50 mya (Meredith *et al.* 2009), so that only eight loci could be identified this way. For locus KoRV-CCTT, one flank was identified in the flank sequence dataset for Pci-SN345, and was used to query low-coverage GSF FLX genomic sequence from Pci-SN404, identifying the other flank in a chromosome without the provirus. For locus KoRV-GCCT, matching 5' and 3' target site duplication sequences were detected after single-flank analyses were conducted (see below); PCR combining a primer from each of the two flanks established (after amplification and sequencing in a chromosome without the provirus) that the 2

flanks corresponded to the same locus. After both flanks for ten KoRV loci were identified, a PCR strategy was utilized to determine whether KoRV was present in an individual koala at a particular locus in both chromosomes, in one chromosome, or in neither of the two chromosomes (Figure 1.1). Three different combinations of primers were used, each combination in a separate PCR reaction for each individual koala (Roca *et al.* 2004). Two of the primer pairs established the presence of the 5' or the 3' flank and LTR, while the other primer pair would amplify only if KoRV was not present at the locus in at least one of the two chromosomes. Using this strategy, the six koalas were screened for insertional polymorphisms across the 10 proviruses (Table 1.2).

The screening involved 6 koalas, ten proviral loci, and two chromosomes for each locus, a total of 120 potential integration sites. Across the 120 sites, a provirus was detected only in 16 cases. In every one of these cases, the provirus was present at a locus in only one of the two chromosomes in an individual. If the progeny koala Pci-SN404 is excluded, since he could have received enKoRVs vertically from either parent, and only the 5 unrelated koalas are considered, then each of these 10 KoRV proviruses was detected in only one koala individual. The lack of shared KoRV proviral loci among unrelated individuals, and the presence of each KoRV provirus in only one of the two chromosomes present in a single individual, suggested that the KoRV proviruses were present at only low frequencies across the koala population, which is consistent with estimates that KoRV only recently entered the koala germ line.

Every northern Australian koala carries many copies of KoRV (Simmons *et al.* 2012). Although the KoRV copy number estimated for northern Australian koalas is 165 copies/cell (Simmons *et al.* 2012), the variance across these koalas is limited (range 139–

199 copies/cell) (Simmons *et al.* 2012). The limited range in copy number may reflect a tendency of random mating to equilibrate the number of enKoRVs per individual within a population. In contrast, across populations, studies of genetic diversity in koalas suggest that gene flow may be limited (Houlden *et al.* 1999). This may be particularly true between koala populations in northern and southern Australia, as the average copy number for KoRV is very low in the south relative to the north, whereas KoRV has been ubiquitous in the north for more than a century (Avila-Arcos *et al.* 2013; Simmons *et al.* 2012; Tarlinton *et al.* 2006). To the degree that gene flow can occur between north and south, this would be expected to eventually equilibrate the copy number of enKoRVs at a level intermediate between those currently found in northern and southern Australian koalas.

In summary, the northern Australian koala population is now marked by a very large number of enKoRV loci, but with each distinct enKoRV at low frequency in the population. Thus only a small proportion of enKoRVs would be shared between individuals, or present in both chromosomes of an individual. Our results suggest that the initial emergence of ERVs involves a massive proliferation of proviruses in the germ lines of one or more populations of the host species. After stabilization, the number of copies of the ERV would be reduced by selection against deleterious integrants; the number of ERV loci would be reduced by drift (with most disappearing but a small proportion becoming fixed), whereas admixture with populations that carry few or no copies of the ERV would lead to dilution and equilibration of ERV copy number.

Figures and Tables

Figure 1.1. Strategy to detect KoRV proviral integration sites in koala genomes.

The genome walking method was modified from that of Reddy et al. (Reddy *et al.* 2008). Using Phi29 DNA polymerase included in the REPLY-g Mini Kit (Qiagen), partially degenerate walker adapters (Table 1.5) were randomly attached across the koala genome (Step 1). Approximately 100 ng of each koala genomic DNA was denatured, following the REPLI-g kit protocols. Four different walker-adapter primers were then attached to each denatured DNA (Reddy *et al.* 2008) (Table 1.5), using a mix that consisted of 10 units of Phi29 DNA polymerase, 1X Phi29 DNA polymerase reaction buffer, 200 μM dNTPs, and 20 μM of each walker-adapter primer. The mixture was incubated at 30°C for 90 minutes to initiate multiple primer extension events, then incubated at 65°C for 10 minutes to inactivate the polymerase. The QIAquick PCR Purification Kit (Qiagen) was used to remove unincorporated walker-adapter primers, following the manufacturer's protocol. The purified DNA fragments with walker-adapter primers were eluted using 40 μl of TLE buffer.

The DNA with walker adapters attached was used to isolate the genomic regions flanking KoRV loci using two amplifications (Step 2). An initial PCR included a primer matching the KoRV LTR with a second primer matching the walker adapters (Table 1.5). Primers used in a second, nested PCR were prepared using the "Roche Genome Sequencer System" (Roche Applied Science, Penzberg, Germany). One primer consisted of three concatenated segments: a GS FLX Titanium adapter "Primer A" segment, a multiplex identifier (MID), and a KoRV specific primer (Table 1.5). The MID used was

Figure 1.1. (cont.)

the same across the four different amplicons of walker adapter but distinctive for each koala individual, and for each run (5' or 3'). The second primer consisted of two concatenated segments: the GS FLX Titanium adapter "Primer B" and a second walker primer (Table 1.5). PCR was conducted using the FastStart High Fidelity PCR System (Roche Applied Science) and the PCR components and algorithm conformed to the manufacturer's protocol. The resulting PCR amplicons were purified using AMPure XP beads (Beckman Coulter, CA USA) with a magnetic particle concentrator. The concentrations of the purified nested PCR amplicons were estimated using a Qubit 2.0 Fluorometer (Life Technologies Corp.) and amplicon sizes, quality, and quantity were measured using an Agilent 2100 Bioanalyzer at the Functional Genomics Unit, Biotechnology Center (Biotech Center) at University of Illinois at Urbana-Champaign (UIUC), USA. Amplicon concentrations were adjusted so that equal amounts would be pooled. The pooled sample was eluted on an agarose gel and separated into 2 size classes, one ~200-400 bp and the other ~400-1000 bp at the High-Throughput Sequencing and Genotyping Unit, Biotech Center at UIUC. Each size class was run separately on $1/16^{th}$ of a PicoTiterPlate (PTP) (1/8 PTP total) on the Roche 454 GS FLX+ platform at the UIUC High-Throughput Sequencing and Genotyping Unit.

Since 5'-LTR and 3'-LTR are nearly identical, half of the PCR amplicons would include the target LTR and host flanking regions (5' flank and 3'flank in Step 2) whereas the other half of the amplicons would inadvertently represent KoRV genes (5' KoRV and 3' KoRV in Step 2). The concentration-adjusted pooled samples were sequenced on the Roche 454 GS FLX+ platform; those sequences that included the host genomic flanking

Figure 1.1. (cont.)

sequences were retrieved and separated by koala using a bioinformatics routine (Step 3). Reads generated by the Roche 454 GS FLX platform were converted into FASTQ format using the Galaxy platform (Giardine *et al.* 2005). The experimentally ligated MID formed part of the sequence read and indicated which koala the sequence originated in, and whether the 5' end of the LTRs or the 3' end of the LTRs was the target. To remove reads matching the KoRV provirus, we used Bowtie2 (Langmead & Salzberg 2012), using the "very sensitive local alignment" preset, to attempt to map all reads to published KoRV sequence AF151794 (Hanger *et al.* 2000). Only reads that did not map to KoRV genes were further considered.

To identify the boundary between the KoRV LTR and the koala flanking genomic sequence, the flanks were mapped onto the published KoRV LTR sequence using Bowtie2 (Langmead & Salzberg 2012), using the "very sensitive local alignment" preset. The LTR sequences proved to be 2 bp shorter at the 5' end and 1 bp shorter for 3' end of the LTRs (total 3 bp shorter) than the published reference sequence (Hanger *et al.* 2000). To carefully determine the boundary between LTR and host genomic flank, the flank sequences were trimmed to include approximately 10 bp of the end of the LTR and 10 bp the koala genomic regions. Each flank sequence was aligned to the Meug_1.1 assembly of the genome of the tammar wallaby (Renfree *et al.* 2011) using BLASTN (Altschul *et al.* 1990) using parameters for short local alignment. Flanks that aligned to more than three scaffolds were removed to reduce the possibility that multiple unique flanks of KoRV might be misidentified as one insertion. We wrote a routine using BioPython (Cock *et al.* 2009) to filter the BLAST results for pairs of 5' and 3' koala genomic flanks.

Figure 1.1. (cont.)

The matched 5’ and 3’ flanks should be aligned within 10 bp due to target site duplication but not overlapping by more than 10 bp, on the same wallaby scaffold, in the proper orientation, to identify koala genomic sequences that corresponded to the 5’ and 3’ host genomic flanks of the same KoRV locus. The trio of sequences (5' of the proviral integration site, 3’ of the proviral integration site, plus the matching wallaby segment) were then re-aligned and visually inspected in the software Sequencher 5.1 (Gene Codes Corp., MI, USA).

Genomic DNA of Pci-SN404 was sequenced on 1/16th of a PicoTiterPlate (PTP) on the Roche 454 GS FLX+ platform (Roche Applied Science) run at the High-Throughput Sequencing and Genotyping Unit, UIUC, and low coverage of genome sequences were available (Ruiz-Rodriguez *et al.* 2014). To identify additional matched flanking sequences on either side of the same proviral locus, all flank sequences were searched against whole genome sequences. For this search, Bowtie2 (version 2.1.0) (Langmead & Salzberg) was run on the Galaxy platform (Giardine *et al.* 2005).

To estimate the number of distinct retroviral integrations from the host flanks sequenced by the Roche 454 GS FLX platform (Table 1.3), the reads were trimmed to only include ca. 50 bp of host genomic flank adjacent to the proviral LTR. We retained only those reads that contained at least 50 bp of host genomic flank and had a base call quality of 99% for every position in the 50 bp. This minimized the possibility of an inflated count due to sequencing errors. For each MID dataset iteration, we used the Megablast algorithm in BLASTN (Altschul *et al.* 1990) to cross-align all filtered reads from the same iteration, and grouped together those reads that were at least 80% similar,

Figure 1.1. (cont.)

with each group of reads counted as a 'distinct' flank sequence. These criteria for grouping the number of distinct flank sequences may have somewhat underestimated the total. For each distinct grouping of flank sequences, the consensus 4 bp at the LTR boundary was taken as the target site duplication for the integration site. The number of proviruses for each koala was estimated as the number of distinct flank sequences detected for 5’ and 3’ flanks separately, and present in at least two of the sequence reads (singletons were removed to minimize potential error). Then for each target site duplication, the number of 5’ and 3’ distinct sequences was compared, and the larger of the two for each target site duplication was used in estimating the total number of reads for each individual koala (Table 1.3). The number of sequencing reads per distinct flank is shown in Figure 1.2.

Figure 1.1. (cont.)

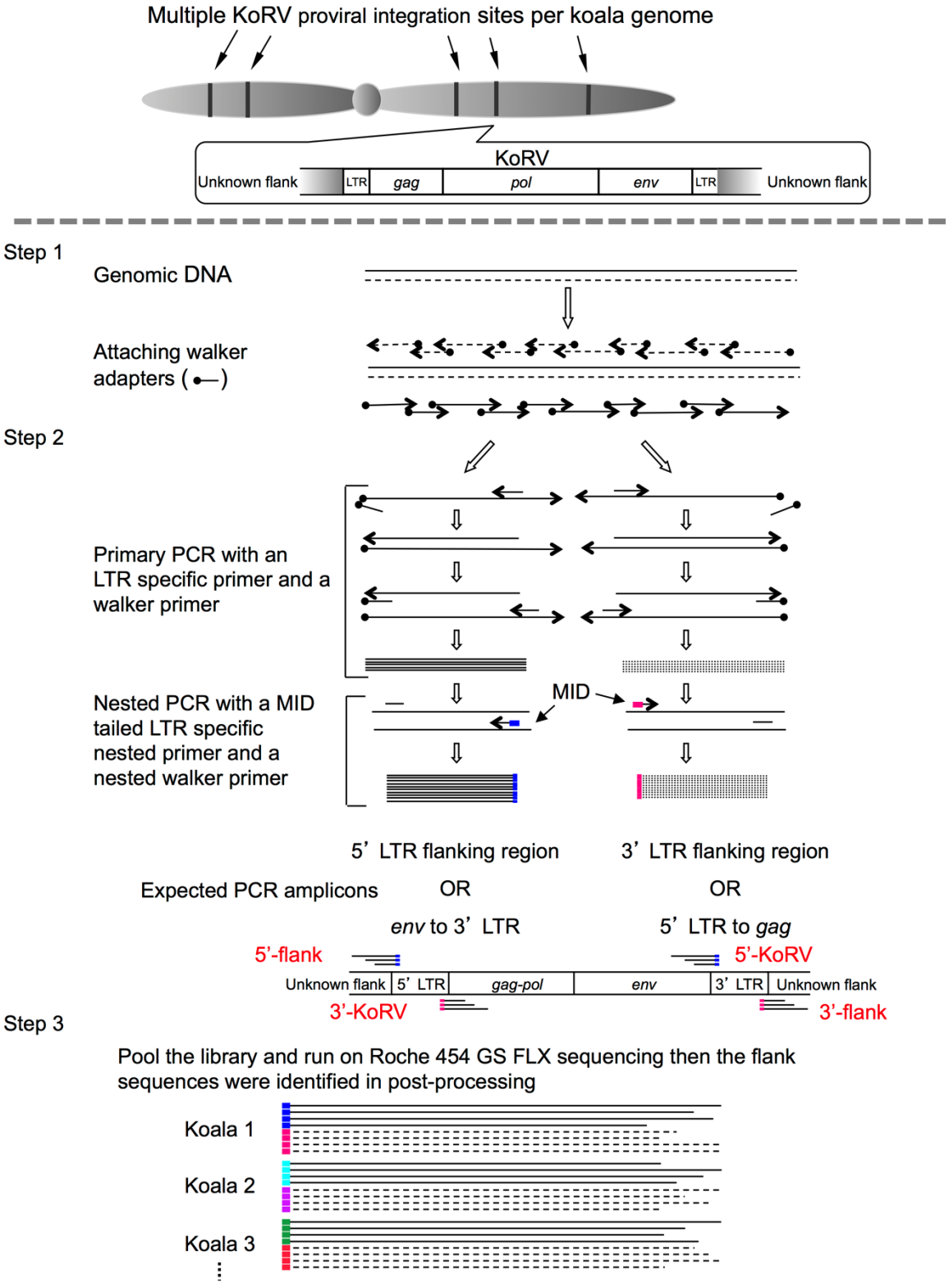

Figure 1.2. Roche 454 GS FLX Platform sequencing reads per distinct KoRV flank sequence, in northern Australian koalas.

Koala genomic flank sequences that were similar were grouped together, after comparing only those reads with 99% base call accuracy in the first 50 bp of the koala genomic flank adjacent to the proviral LTR. The *x*-axis shows coverage, or the number of sequence reads corresponding to each distinct flank sequence (bins are of equal range except for flanks with >500 reads). The *y*-axis indicates, for each range of coverage, the number of distinct proviral flanks that had that level of coverage. Panels represent results for different koalas: (A) Pci-SN404, progeny; (B) Pci-SN248, sire; and (C) Pci-SN345, dam. The majority of flanks were covered by a low or moderate number of sequence reads, and the genome walking method isolated more than just a few proviral loci.

Figure 1.2. (cont.)

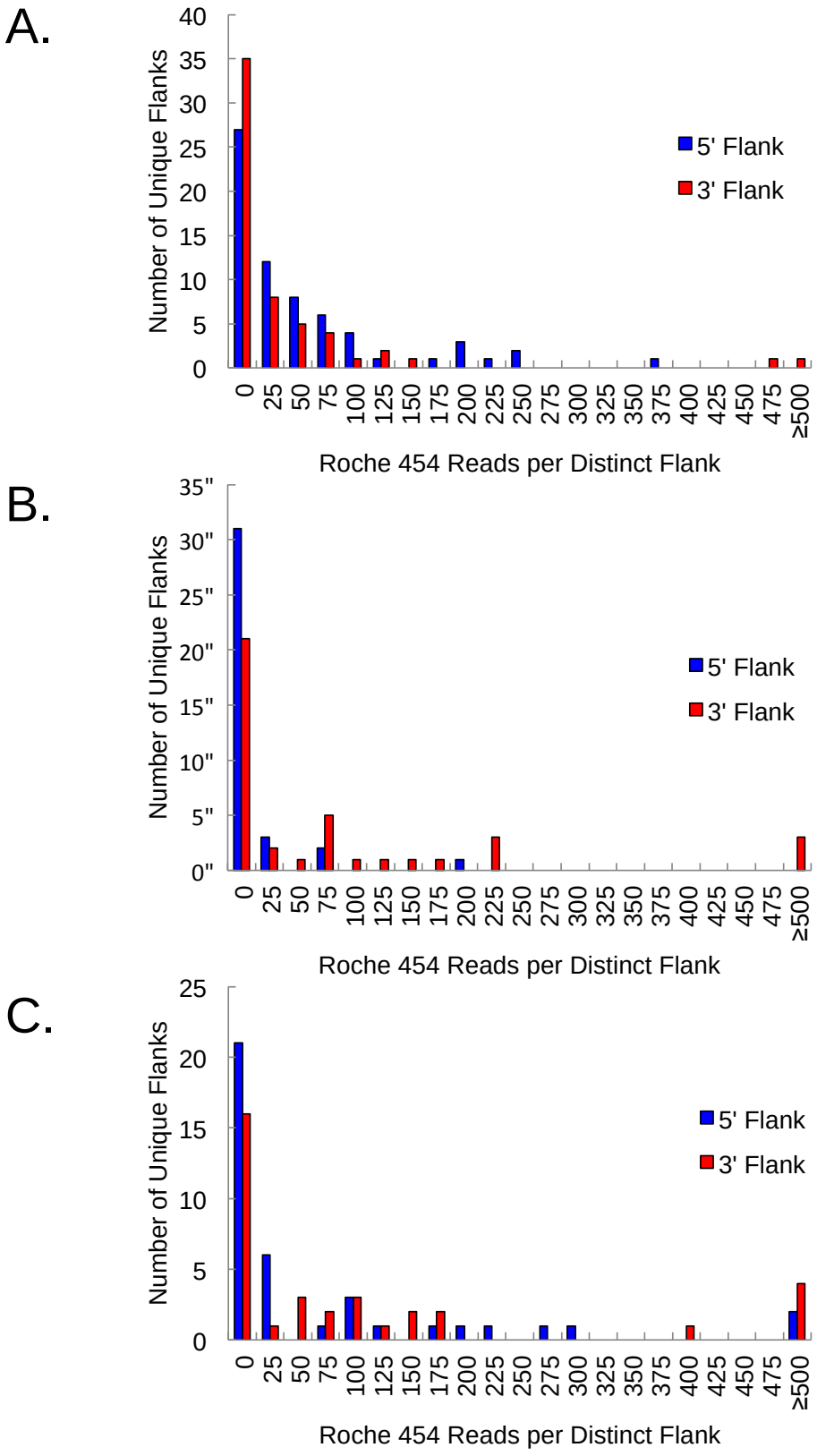

Table 1.1a. Northern Australian koala triad description and KoRV flank statistics.

Sample ID	Pci-SN404		Pci-SN248		Pci-SN345	
Genealogy	Son of SN248/345		Sire of SN404		Dam of SN404	
Birth Date	October 31, 2006		September 29,1998		February 18, 2003	
Place sample collected	Columbus Zoo		San Diego Zoo		San Diego Zoo	
Year sample collected	2010		2010		2010	
5' or 3' LTR and flank	5'	3'	5'	3'	5'	3'
Total 454 sequencing reads	21556	19554	7869	15633	18640	21223
KoRV LTR plus flank reads	8391	7212	2465	7263	7392	10167
Distinct flanks	76	68	43	45	43	39

Table 1.1b. Southern Australian koala sample descriptions and KoRV flank statistics.

Sample ID	Pci-157		Pci-106		Pci-187	
Place sample collected	Stony Rises		Brisbane Ranges		Kangaroo Island	
Year sample collected	1991		1991		1991	
5' or 3' LTR and flank	5'	3'	5'	3'	5'	3'
Total 454 sequencing reads	14608	9804	1275	3987	1041	1240
KoRV LTR plus flank reads	6134	3907	9	1198	1	1
Distinct flanks[†]	11	8	0	10	0	0

Table 1.2. Insertional polymorphisms of KoRV

	Northern koala triad, Pci-			Southern koala triad, Pci-		
Provirus	SN404	SN248	SN345	157	106	187
KoRV-ACAT	+/-	-/-	+/-	-/-	-/-	-/-
KoRV-CTAG	+/-	-/-	+/-	-/-	-/-	-/-
KoRV-AAAAG	+/-	-/-	+/-	-/-	-/-	-/-
KoRV-AAGT	+/-	-/-	+/-	-/-	-/-	-/-
KoRV-CCTT	-/-	-/-	+/-	-/-	-/-	-/-
KoRV-AAAG	+/-	-/-	+/-	-/-	-/-	-/-
KoRV-CCCC	+/-	-/-	+/-	-/-	-/-	-/-
KoRV-GCCT	+/-	-/-	+/-	-/-	-/-	-/-
KoRV-GTAC	-/-	+/-	-/-	-/-	-/-	-/-
KoRV-ACTT	+/-	-/-	+/-	NA	NA	NA

+/+, provirus present on both chromosome homologs.
+/-, provirus present on only one of the two homologs.
-/-, neither chromosome had a provirus at the locus.
KoRVs were first identified in Pci-SN404 except for CCTT (Pci-SN345) and GTAC (Pci-SN248).
Boxes enclose proviral loci with identical LTR sequence.
For KoRV-ACTT there was no amplification (NA) in southern Australian koalas.

Table 1.3. Number of unique proviruses per 4 bp target site duplication.

	Northern koalas									Southern koalas								
	Pci-SN404 Columbus zoo Progeny			Pci-SN248 San Diego zoo Sire			Pci-SN345 San Diego zoo Dam			Pci-157 Stony Rises NA			Pci-106 Brisbane Ranges NA			Pci-187 Kangaroo Island NA		
	5' flank	3' flank	*	5' flank	3' flank	*	5' flank	3' flank	*	5' flank	3' flank	*	5' flank	3' flank	*	5' flank	3' flank	*
Total	**76**	**68**	**108**	**43**	**45**	**73**	**43**	**39**	**69**	**11**	**8**	**16**	**0**	**10**	**10**	**0**	**0**	**0**
AAAC	2	1	2	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
AAAG	2	2	2	0	1	1	0	0	0	6	3	6	0	0	0	0	0	0
AAAT	2	0	2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
AAGC	1	1	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
AAGG	2	1	2	2	0	2	3	0	3	0	0	0	0	0	0	0	0	0
AAGT	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AATG	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ACAA	1	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
ACAG	1	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
ACAT	1	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
ACTA	0	0	0	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0
ACTG	2	1	2	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
ACTT	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
AGAC	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
AGAT	0	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
AGGC	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AGGG	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AGGT	0	0	0	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0
AGTC	2	1	2	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
AGTT	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ATAC	2	1	2	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0

Table 1.3 (cont.) Number of unique proviruses per 4 bp target site duplication.

	Northern koalas									Southern koalas								
	Pci-SN404			Pci-SN248			Pci-SN345			Pci-157			Pci-106			Pci-187		
	Columbus zoo			San Diego zoo			San Diego zoo			Stony Rises			Brisbane Ranges			Kangaroo Island		
	Progeny			Sire			Dam			NA			NA			NA		
	5' flank	3' flank	*	5' flank	3' flank	*	5' flank	3' flank	*	5' flank	3' flank	*	5' flank	3' flank	*	5' flank	3' flank	*
Total	**76**	**68**	**108**	**43**	**45**	**73**	**43**	**39**	**69**	**11**	**8**	**16**	**0**	**10**	**10**	**0**	**0**	**0**
ATAC	2	1	2	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0
ATAG	2	4	4	1	2	2	0	0	0	0	0	0	0	0	0	0	0	0
ATAT	1	2	2	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
ATCA	1	2	2	0	0	0	4	6	6	0	0	0	0	0	0	0	0	0
ATCC	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ATGA	0	0	0	4	2	4	0	0	0	0	0	0	0	0	0	0	0	0
ATGC	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
ATGG	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
ATTC	0	1	1	0	0	0	0	2	2	0	0	0	0	0	0	0	0	0
ATTT	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CAAC	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CACT	1	0	1	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0
CAGC	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
CATA	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
CATC	0	0	0	0	0	0	2	4	4	0	0	0	0	0	0	0	0	0
CCAC	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CCAG	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
CCAT	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CCCC	4	2	4	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
CCCT	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0

Table 1.3 (cont.) Number of unique proviruses per 4 bp target site duplication.

	Northern koalas									Southern koalas								
	Pci-SN404 Columbus zoo Progeny			Pci-SN248 San Diego zoo Sire			Pci-SN345 San Diego zoo Dam			Pci-157 Stony Rises NA			Pci-106 Brisbane Ranges NA			Pci-187 Kangaroo Island NA		
	5' flank	3' flank	*	5' flank	3' flank	*	5' flank	3' flank	*	5' flank	3' flank	*	5' flank	3' flank	*	5' flank	3' flank	*
CCTT	0	0	0	0	0	0	2	0	2	0	0	0	0	0	0	0	0	0
CTAA	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
CTAC	0	4	4	2	1	2	0	0	0	0	0	0	0	0	0	0	0	0
CTAG	3	2	3	0	0	0	0	2	2	0	0	0	0	0	0	0	0	0
CTAT	3	2	3	5	1	5	0	0	0	0	1	1	0	0	0	0	0	0
CTCC	1	0	1	0	1	1	0	0	0	0	0	0	0	7	7	0	0	0
CTGA	0	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
CTTA	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CTTC	0	1	1	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0
GAAC	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
GAAG	3	1	3	3	2	3	0	0	0	0	0	0	0	0	0	0	0	0
GAAT	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
GACC	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GAGC	1	0	1	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0
GATA	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
GATC	3	2	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GATG	0	0	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0
GCAC	1	1	1	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0
GCCG	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
GCCT	2	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GCTT	1	1	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
GGAA	2	0	2	0	0	0	2	0	2	0	0	0	0	0	0	0	0	0

Table 1.3 (cont.) Number of unique proviruses per 4 bp target site duplication.

	Northern koalas									Southern koalas								
	Pci-SN404			Pci-SN248			Pci-SN345			Pci-157			Pci-106			Pci-187		
	Columbus zoo			San Diego zoo			San Diego zoo			Stony Rises			Brisbane Ranges			Kangaroo Island		
	Progeny			Sire			Dam			NA			NA			NA		
	5' flank	3' flank	*	5' flank	3' flank	*	5' flank	3' flank	*	5' flank	3' flank	*	5' flank	3' flank	*	5' flank	3' flank	*
GGAC	1	1	1	2	3	3	0	0	0	0	0	0	0	0	0	0	0	0
GGAG	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
GGAT	1	3	3	0	0	0	2	2	2	0	0	0	0	0	0	0	0	0
GGCC	0	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
GGGC	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
GGTA	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
GGTC	0	0	0	0	0	0	0	2	2	0	0	0	0	0	0	0	0	0
GTAC	1	2	2	2	4	4	1	0	1	0	0	0	0	0	0	0	0	0
GTAG	1	1	1	2	0	2	2	0	2	0	0	0	0	0	0	0	0	0
GTAT	2	1	2	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
GTCT	0	0	0	0	3	3	0	0	0	0	0	0	0	0	0	0	0	0
GTGC	1	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
GTGG	0	1	1	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0
GTTG	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
GTTT	0	1	1	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0
TAAT	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TAGA	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
TAGG	0	0	0	1	0	1	3	0	3	0	0	0	0	0	0	0	0	0
TATG	0	1	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
TCAT	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TCTC	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TGAG	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0

Table 1.3 (cont.) Number of unique proviruses per 4 bp target site duplication.

	Northern koalas									Southern koalas								
	Pci-SN404			Pci-SN248			Pci-SN345			Pci-157			Pci-106			Pci-187		
	Columbus zoo			San Diego zoo			San Diego zoo			Stony Rises			Brisbane Ranges			Kangaroo Island		
	Progeny			Sire			Dam			NA			NA			NA		
	5' flank	3' flank	*	5' flank	3' flank	*	5' flank	3' flank	*	5' flank	3' flank	*	5' flank	3' flank	*	5' flank	3' flank	*
TGCA	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TGCT	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
TGGC	2	0	2	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
TGGT	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
TGTT	0	4	4	0	6	6	0	4	4	0	1	1	0	2	2	0	0	0
TTAC	1	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
TTAG	0	1	1	0	0	0	0	1	1	5	0	5	0	0	0	0	0	0
TTAT	2	2	2	1	2	2	1	0	1	0	0	0	0	0	0	0	0	0
TTCC	1	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
TTCT	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
TTGC	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
TTTC	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
TTTG	0	2	2	0	2	2	1	0	1	0	3	3	0	1	1	0	0	0

Reads containing similar koala flanks were grouped together using the first 50 bp of the koala genomic region from the viral LTR and defined to adjacent to the same proviral locus. Counts indicate the number of unique proviral loci that has the same target site duplication but different flank sequences that are supported by at least two reads. The column denoted by * is an estimate of unique integrants for that individuals, which is the greater of the 5' and 3' flanks.

Table 1.4. Total reads (non-unique) detected by Roche 454 GS FLX platform of each proviral flank sequence.

	Pci-SN404 Progeny		Pci-SN248 Sire		Pci-SN345 Dam	
Proviral locus	5' flank	3' flank	5' flank	3' flank	5' flank	3' flank
KoRV-AAAAG	4	52	0	0	0	0
KoRV-AAAG	1	0	0	0	0	0
KoRV-AAGT	3	13	0	0	0	0
KoRV-ACAT	35	0	0	0	0	0
KoRV-ACTT	85	0	0	0	0	0
KoRV-CCCC	97	0	0	0	0	0
KoRV-CCTT	0	0	0	0	116	0
KoRV-CTAG	0	68	0	0	0	0
KoRV-GCCT	108	0	0	0	0	0
KoRV-GTAC	0	0	10	0	0	0
KoRV-5-AAGG	3	–	7	–	0	–
KoRV-5-AGTC	396	–	0	–	0	–
KoRV-5-ATAG	29	–	0	–	0	–
KoRV-5-ATGG	1	–	0	–	221	–
KoRV-5-CAAC	6	–	0	–	0	–
KoRV-5-CCCC	53	–	0	–	0	–
KoRV-5-CTAG	30	–	0	–	0	–
KoRV-5-CTAT	1	–	0	–	0	–
KoRV-5-GAAG	30	–	0	–	0	–
KoRV-5-GAGC	5	–	0	–	0	–
KoRV-5-GCTT	84	–	0	–	15	–
KoRV-5-GTGC	139	–	9	–	0	–
KoRV-5-GTTG	68	–	0	–	0	–
KoRV-5-TCAT	66	–	0	–	0	–
KoRV-5-TGCA	241	–	0	–	0	–
KoRV-5-TTAC	0	–	0	–	0	–
KoRV-5-TTAT	72	–	0	–	0	–
KoRV-5-TTCC	200	–	0	–	205	–
KoRV-3-AAAT	–	92	–	2	–	0
KoRV-3-AGAT	–	101	–	0	–	190
KoRV-3-AGGC	–	2	–	0	–	0
KoRV-3-ATAG	–	498	–	0	–	0
KoRV-3-ATCA	–	0	–	0	–	0
KoRV-3-CTTC	–	0	–	0	–	0
KoRV-3-GATC	–	136	–	0	–	0
KoRV-3-GGAT	–	2	–	0	–	136

Table 1.4 (cont.) Total reads (non-unique) detected by Roche 454 GS FLX platform of each proviral flank sequence.

KoRV-3-GGCC	–	8	–	0	–	35
KoRV-3-GGTA	–	0	–	0	–	0
KoRV-3-TTAT	–	85	–	0	–	0

The first ten rows of this table shows 10 enKoRV sequences with matched 5' and 3' flanks. The remainder of the table are enKoRV loci with only one flank identified in the Roche 454 GS FLX sequencing reads.
KoRV-3-GGTA and KoRV-3-CTTC were not counted, due to tandem repeats of homonucleotides in the sequences which will distort the alignment algorithm used for finding similar reads.

Table 1.5. The nucleotide sequences of walker adapters and of KoRV LTR primers used to identify proviral integration sites.

Name	Sequence	Note	Primer position in KoRV
Walker adapters and primers published by Reddy et al. 2008			
Walker adapter 1	GTGAGCGCGCGTAATACGACTCACTATAGGGNNNNATGC		
Walker adapter 2	GTGAGCGCGCGTAATACGACTCACTATAGGGNNNNGATC		
Walker adapter 3	GTGAGCGCGCGTAATACGACTCACTATAGGGNNNNTAGC		
Walker adapter 4	GTGAGCGCGCGTAATACGACTCACTATAGGGNNNNCTAG		
Walker primer 1	GTGAGCGCGCGTAATACGA	Used for primary PCR	
Walker primer 2*	GTAATACGACTCACTATAGGG	Used for nested PCR	
KoRV specific primers to detect 5' flanks			
5'LTR-R21	CCTTGTTTTTCTTGCTCTGACC	Used for primary PCR	5' LTR (b in Figure 1.1A)
5'LTR-R22†	CACCTGTCCCTAAACCTTGG	Used for nested PCR	5' LTR (b in Figure 1.1A)
KoRV specific primers to detect 3' flanks			
3'LTR-F	ATTTGCATCCGGAGTTGTGT	Used for primary PCR	3' LTR (c in Figure 1.1A)
3'LTR-F2†	AGTTGTGTTCGCGTTGATCC	Used for nested PCR	3' LTR (c in Figure 1.1A)

*: Walker primer 2 had the following sequences concatenated in a single oligonucleotide primer: (1) the GS FLX Titanium adapter Primer A (CCATCTCATCCCTGCGTGTCTCCGACTCAG), (2) a multiplex identifier (MID), and (3) the sequence shown above.
†: The 5'LTR-R22 primer had the following sequences concatenated in a single oligonucleotide primer: (1) GS FLX Titanium adapter Primer B (CCTATCCCCTGTGTGCCTTGGCAGTCTCAG) and (2) the sequence shown above.
KoRV specific primers were also used to identify enKoRV in combination with primers listed in Table 1.7.

Table 1.6. Primers used for loci at which genomic flanks on both sides of the KoRV provirus were identified.

Locus name	Primer name	Oligonucleotide sequence	Primer position
	Flank specific primers		
KoRV-AAAA	KoRV_flankC_F1	TCCGTATCCCATATGCTGTG	5' flank (a)
	KoRV_flankC_F2	ATCCCCTTCTCCTCCAAAGA	5' flank (a)
	KoRV_flankC_R1	GGAAGGCCAGCTAGGTTAGG	3' flank (d)
	KoRV_flankC_R2	CCCATGGGTTTTTCTCAGTT	3' flank (d)
KoRV-AAAG	KoRV_flankA_F1	CTAAGCTTGTCCCCGGAACT	5' flank (a)
	KoRV_flankA_F2	CTAAGCTTGTCCCCCTGCTT	5' flank (a)
	KoRV_flankA_R1	CTGCAGCAAAATCCCAGAAT	3' flank (d)
	KoRV_flankA_R2	GCAGCAAAATCCCAGAATTG	3' flank (d)
KoRV-AAGT	KoRV_flankD_F3	TTCCGAACTTGGGTAAGCAT	5' flank (a)
	KoRV_flankD_F4	CTTCCGAACTTGGGTAAGCAT	5' flank (a)
	KoRV_flankD_R3	ACCAAATTATGAAAAGTTGCTTGA	3' flank (d)
	KoRV_flankD_R4	CAAATTATGAAAAGTTGCTTGACA	3' flank (d)
KoRV-ACAT	KoRV_flankE_F1	AGTTTTCCCAGTCACACAGGA	5' flank (a)
	KoRV_flankE_F2	TTTTCCCAGTCACACAGGACT	5' flank (a)
	KoRV_flankE_R1	TGGTTGTATTGATTTGTATGATTCC	3' flank (d)
	KoRV_flankE_R2	TTGTGAGCTCTCTGATTGGTTC	3' flank (d)
	KoRV_flankE_R3	GTGAGCTCTCTGATTGGTTCAA	3' flank (d)
KoRV-ACTT	KoRV_flankF_F5	AAGGGACCTTAGAAACCACGTA	5' flank (a)
	KoRV_flankF_F6	GGGACCTTAGAAACCACGTAGC	5' flank (a)
	KoRV_flankF_R3	GGGTGGTACATGGTTTCTTTTC	3' flank (d)
KoRV-CCCC	KoRV_flankH_F1	TGTTCCAGGGAAGGAAATGA	5' flank (a)
	KoRV_flankH_F2	TTTCATTGTTCCAGGGAAGG	5' flank (a)
	KoRV_flankH_R1	AAGGAGCCCTGGGTGTTT	3' flank (d)
	KoRV_flankH_R2	CAAGGAGCCCTGGGTGTT	3' flank (d)
KoRV-CCTT	KoRV_flankJ_F1	CTACCTGAGTCCCTTCCCAAT	5' flank (a)
	KoRV_flankJ_F2	CTGAGTCCCTTCCCAATTTT	5' flank (a)
	KoRV_flankJ_R1	GGACTTTCCAGCAGAGTTCTATATG	3' flank (d)

Table 1.6 (cont.) Primers used for loci at which genomic flanks on both sides of the KoRV provirus were identified.

Locus name	Primer name	Oligonucleotide sequence	Primer position
KoRV-CTAG	KoRV_flankB_F1	TCAGCCATTAAATGTCAAGCA	5' flank (a)
	KoRV_flankB_F2	CAGCCATTAAATGTCAAGCAGA	5' flank (a)
	KoRV_flankB_F3	CAATTGGGAACTAGGATGAAATG	5' flank (a)
	KoRV_flankB_F4	TTGGGAACTAGGATGAAATGAAC	5' flank (a)
	KoRV_flankB_R1	TTCCAAGCGTTGTTCATTTG	3' flank (d)
	KoRV_flankB_R2	TGTTCATTTGCTCCCTCTCA	3' flank (d)
KoRV-GCCT	N5-18-F	ATAGAGCATTGGCCTTGGTG	5' flank (a)
	N5-18-F2	CTAGGTGGCACGGTGGATAG	5' flank (a)
	N3-3-R	ACACGAACCATCCATCCATT	3' flank (d)
	N3-3-R2	GAACCATCCATCCATTGCTT	3' flank (d)
KoRV-GTAC	KoRV_flankI_F1	AGTCAAACGGAATTGTAATCTGA	5' flank (a)
	KoRV_flankI_F2	TTAGTCAAACGGAATTGTAATCTGA	5' flank (a)
	KoRV_flankI_F3	CTTAGTCAAACGGAATTGTAATCTGA	5' flank (a)
	KoRV_flankI_F4	TAGTCAAACGGAATTGTAATCTGA	5' flank (a)
	KoRV_flankI_R1	GACCAGGATGTAGGGCAGAC	3' flank (d)
	KoRV_flankI_R2	CCAGGATGTAGGGCAGACAA	3' flank (d)
KoRV specific primers			
KoRV-*gag-pol*	PCI-KoRV-R1.2	AATCTCAGATCCCGGACGA	*gag* (b2)
KoRV-*gag-pol*	PCI-KoRV-R1.3	GGTCCTTGGGTGGGAATCT	*gag* (b2)
KoRV-*env*	PCI-KoRV-F29	CAGACCCTAGACAACGAGGA	*env* (c2)
KoRV-*env*	PCI-KoRV-F29.2	TTCTGGTTCTCAGGCACAAG	*env* (c2)

M13 forward and reverse sequences that were attached to each forward and reverse primer, respectively, are not shown here. Primers were combined as in Figure 1.1. The letter within parentheses accords with the primer position shown in Figure 1.1. Depending on the sequence quality of the Roche 454 GS FLX+ platform and to avoid failing to amplify the target region, more than one primer was designed for each proviral locus.

Table 1.7. Primers used to screen for endogenous KoRV loci.

Locus name	Primer name	Sequence
	5' flank	
KoRV-5-AAGG	N5-15-F	CAATGGTTCAAAGTATGCCTAGTG
KoRV-5-AGTC	N5-4-F	CCTGGGCCCTCTTTTCTCTA
KoRV-5-ATAG	N5-14-F	CATGACCTCTGGTTGTGATGA
KoRV-5-ATGG	KoRV_flankG_F1*	TGTTCACACTGTTCATGCAAAT
KoRV-5-ATGG	KoRV_flankG_F2*	TCACACTGTTCATGCAAATAGC
KoRV-5-ATGG	KoRV_flankG_F3*	TTCTACCCATATTCCTCTCATTCC
KoRV-5-CAAC	N5-23-F	TGCAAGCTACTCACTTTGGAGA
KoRV-5-CCCC	N5-13-F	GGGTTCAAATTGTGCTTCCA
KoRV-5-CTAG	N5-8-F	GCATCCGGTAACTCTGAGGA
KoRV-5-CTAT	N5-17-F	CAAGACAGGAATGGATTTTATGT
KoRV-5-GTTG	N5-12-F	AACTGCATTGAGCCAGGTTT
KoRV-5-GAAG	N5-19-F	AGTCACAGGGCTGTCAATG
KoRV-5-GAGC	N5-24-F	AACATGCTGTTCTTTGAATTGG
KoRV-5-GCTT	N5-11-F	TTGACCTGGACAAGAGAAGACT
KoRV-5-GTGC	N5-5-F	TGATAGAGCTCCAGCCTTGG
KoRV-5-TCAT	N5-21-F	AGGCATCCCTCATTCATTTG
KoRV-5-TGCA	N5-1-F	CGCTTATTGAGAGTGTAGTGCTTT
KoRV-5-TTAC	N5-7-F	TTTGTGAAGCTCAAAGGAGAGA
KoRV-5-TTAT	N5-20-F	TTACAAGGTGAGAACATTGTTTAAGT
KoRV-5-TTCC	N5-2-F	GGATTCAAATCCTGCCTTCA
	3' flank	
KoRV-3-AAAT	N3-4-R	GGGTACTTGACTTAAAATCAGGAAGT
KoRV-3-AGAT	N3-8-R	AACCCCAAAATCACTTTGTCC
KoRV-3-AGGC	N3-17-R	GGATTGTTCTGATGATCACTGC
KoRV-3-ATAG	N3-1-R	GAATGCCACTTTGATGCAGA
KoRV-3-ATCA	N3-22-R	TTGCCTCTGCAGAACAAATAG
KoRV-3-CTTC	N3-21-R	TGTGAATTGCAGCTTTGGAG
KoRV-3-GATC	N3-13-R	GGGGAGGGAATAATGTCCAA
KoRV-3-GGAT	N3-6-R	AAGCACCATTCAAGACCATTG
KoRV-3-GGCC	N3-2-R	CACAATGGCCTCAGCTCTTT
KoRV-3-GGTA	N3-9-R	CTGAAGTCAACAGGGAAGAGC
KoRV-3-TTAT	N3-5-R	GGCTCTAAGGTGGAGAACACC

Primers were designed based on Pci-SN404 flank sequences.
Each 5' flank primer was paired with primer 5'LTR-R21 or 5'LTR-R2 for PCR.
Each 3' flank primer was paired with primer 3'LTR-F or 3'LTR-F2 for PCR.
M13 forward and reverse sequences that were attached to each forward or reverse primer, respectively, are not shown here.
* These primers target the same provirus, KoRV-5-ATGG.

CHAPTER 2: POLYMORPHIC MICROSATELLITE MARKERS IDENTIFIED USING NEXT GENERATION SEQUENCING

Introduction

Microsatellites are tandem repeats, often consisting of 1-6 nucleotides, found in the nuclear genomes of nearly every species (Chambers & MacAvoy 2000). The number of repeats at a particular locus can be highly variable as a consequence of mutations due to strand slippage during replication (Levinson & Gutman 1987). When microsatellites, also known as short tandem repeats (STRs), are comprised of di-, tetra- or penta-nucleotide motifs, tandem repeat number variation would cause frame shifts, thus making them only likely to occur in non-coding regions. Thus most STRs identified and implemented in genetic studies are within selectively neutral loci (Selkoe & Toonen 2006). Since microsatellites are under Mendelian inheritance and are often polymorphic within a population (Okello *et al.* 2005), they can be useful for wildlife management by providing valuable information for determining genetic diversity and population structure within and among populations (Hedges *et al.* 2013); estimating population sizes (censusing) (Eggert *et al.* 2003); assessing population viability (DeSalle & Amato 2004); elucidating historical and contemporary flow patterns and mating systems (Thitaram *et al.* 2008); and identifying the population of origin of illegal wildlife products (Wasser *et al.* 2015).

Many popular programs use sequencing reads to identify and design PCR primers for microsatellite loci including Msatcommander (Faircloth 2008), Msatfinder (Thurston & Field 2005), RepeatMasker (Smit *et al.* 2013-2015) and SciRoKo (Kofler *et al.* 2007). However, *existing programs do not automate screening for putatively polymorphic loci.*

Thus this step requires for primers to be screened in the laboratory across multiple samples in order for loci that are polymorphic to be identified. Existing programs also do not identify microsatellite loci present in repetitive elements that need to be avoided for population analyses. Finally, in our experience, most existing (Tarlinton *et al.* 2006) programs are not capable of handling the large quantity of reads produced by next generation sequencing platforms, such as Illumina. We have therefore developed a program that includes the functionality of existing software and addresses each of the identified shortcomings of existing software, with the objective of generating polymorphic STR markers while limiting the amount of costly and time consuming laboratory experiments required to establish that such markers work effectively and are polymorphic. The latter would be a major concern for species or populations with low levels of genetic diversity (Driscoll *et al.* 2002). In some genetically depauperate species hundreds of monomorphic loci would have to be screened when using traditional methods, in order to find a set of markers that were polymorphic. Such screening can be costly in terms of time, funds and limited DNA samples.

We developed the novel software POLYMSAT to use whole genome shotgun sequencing reads to identify polymorphic microsatellite loci. We integrated Primer3 (Rozen & Skaletsky 2000; Untergasser *et al.* 2012) to automate primer design for loci identified as polymorphic. POLYMSAT also performs *in silico* PCR to identify loci with polymorphic alleles, and to eliminate undesirable primer sets from downstream application, such as those that contain a SNP in their binding site and those that may amplify multiple loci.

Microsatellite marker design often needs to balance two opposing mandates: To allow for amplification in degraded DNA, such as from fecal or ancient samples, amplicons sizes should be minimized. Markers should also be designed such that resulting amplicons contain sufficient unique gene flank to allow sequencing to confirm the locus. By filtering through millions of NGS-reads, POLYMSAT helps identify microsatellite loci that fulfill this dual mandate.

We developed a companion program, PRIMERAWESOME, which generates interactive reports for *in silico* PCR (Figure 2.1). PRIMERAWESOME aggregates primer sets by the genomic region they amplify, by determining which primer pairs will amplify similar sequences. PRIMERAWESOME creates a multiple alignment for the expected amplicons of each of the primer groups, which are visualized with the primer binding sites and base call scores. We believe that these reports will help leverage the expertise of researchers, allowing them to review primers sets prior to time-consuming lab work. PRIMERAWESOME reports are automatically generated with POLYMSAT results. Users can input previously designed oligonucleotide primer sequences, to allow PRIMERAWESOME to identify new microsatellite loci. PRIMERAWESOME reports can be generated in POLYMSAT as a standalone feature, using existing primers and without additional primer design. PRIMERAWESOME can also be used to reconcile primers designed using two separate programs or reconcile an existing set of primers with newly generated primers. Since it groups primers by the sequences of their target amplicons, primers with different sequences that amplify the same locus are easily identified.

To test our novel software, we sought to develop STR markers for the Sumatran rhinoceros, *Dicerorhinus sumatrensis*, a critically endangered species with an estimated

current population size between 140 and 240 individuals with a decreasing trend (Emslie *et al.* 2013). They reside in dense Indonesian forests, making traditional aerial surveys impractical; however, genetic surveys of DNA from dung using microsatellite loci could be used to provide an accurate census estimate and provide additional information on population dynamics (Kohn *et al.* 1999; Taberlet *et al.* 1999).

Methods

We wrote POLYMSAT using C and Python to develop a computationally efficient and user-friendly microsatellite design tool (Figure 2.2). POLYMSAT uses FASTQ or FASTA files for input, which need to contain sequences from the DNA of only one individual. We designed this program to use paired-end reads from the Illumina MiSeq platform and it is suitable for any sequencing platform with comparable base-call confidence and read lengths of at least 150 bp. POLYMSAT must identify at least two reads containing the same microsatellite locus with a different repeat count to label a locus as polymorphic. POLYMSAT can detect heterozygous loci if only one individual is provided, however, sequences from multiple individuals will increase the chances for detection of polymorphic loci. Combined sequencing coverage of all individuals should be at least 2X for POLYMSAT to increase the probability that multiple alleles at the same locus area sequenced. However, POLYMSAT will also accept lower sequencing depth.

SCAN FOR MICROSATELLITE LOCI

POLYMSAT parses FASTA or multiple varieties of FASTQ files and then passes each read into a high-speed microsatellite filter (Figure 2.3). The high speed filter, written in C, iteratively scans the sequence for each motif size of interest. PolyMsat searches for 2-7bp motifs with at least 4 tandem repeats. For motif size *M* and minimum tandem repeats *N*, the read is divided into non-overlapping windows of size *M*. Each window is converted into a single integer that uniquely represents the combination of nucleotides contained. If *N*-1 adjacent windows have the same integer, the read is deemed to potentially include a short tandem repeat.

Since the frame of the sliding windows may not align with the frame of the tandem repeat, the putative repeat region is rescanned in every possible frame. If a repeat of ≥N is found, the extent of the microsatellite array is the largest repeat region found in the offset window scan. Since this strategy will not correctly label motifs comprised of shorter repeats (*e.g.* "TATA" may be incorrectly labeled as a tetranucleotide motif, when it is actually the dinucleotide "TA"), each motif is then verified using a dictionary of true motifs for the repeat size. Once verified, repeat alleles are inserted into a SQLite database, which tracks them through the remainder of the process. We empirically found this method for identifying microsatellite repeats to be faster than scanning using regular expressions, scanning for a list of microsatellites or scanning for repeats using sliding windows without first converting the sequences into integers.

DESIGN AND VALIDATE PRIMERS

POLYMSAT recursively joins adjacent repeat motifs, within a user-specified distance, to create complex microsatellite loci. Using Primer3-Py, a Python interface for Primer3 (Rozen & Skaletsky 2000; Untergasser *et al.* 2012), primer design is attempted at each locus (Figure 2.1, lower panel). Primer design is also attempted on constituent simple microsatellites in complex loci. Although these loci likely lack sufficiently long flanks for primer design, the few that have adequate flanks may be useful for designing for short amplicons. When multiple primer sets meet specifications for a locus, the companion program, PRIMERAWESOME, will display all primer sets together. By default, POLYMSAT requires an 18bp internal flank between the primer binding site and the microsatellite motif, so PCR products could be sequenced to verify the identity of the locus. Primer pairs reported for each repeat region are inserted into the SQLite database.

We developed a custom *in-silico* PCR routine to test designed primers against NGS reads and/or assembled genomes, as many primers will undesirably amplify at multiple genomic loci. Primer binding sites are predicted separately for 5' and for 3' primers; Primers that match genomic loci with up to 2 SNP variants are identified and removed from consideration. Primers are disqualified if either primer has more than one binding site per locus or if they bind any locus with an unexpected orientation.

Some microsatellites may fall within repetitive genomic regions, such as transposable elements, and primers for these may simultaneously amplify multiple loci. We therefore developed a routine to identify such loci. Primers were recursively grouped when their list of predicted amplicons share at least one NGS read. The NGS reads within each group were aligned to each other using BLAST+ after the microsatellite regions were masked from alignment (Camacho *et al.* 2009). If any pairwise alignment did not

include the start or end of either read, or sequence identity was below 95%, the reads were considered to be from repetitive regions and all primers in the group are disqualified. We empirically found that the cutoff of 95% successfully removes sufficient repetitive regions without overzealously eliminating candidate loci.

We developed a companion program, PRIMERAWESOME, using HTML, CSS and JavaScript, to provide interactive primer reports for visualization allowing for the screening of candidate microsatellite loci. PRIMERAWESOME reports are generated automatically by POLYMSAT. They can also be generated as a stand-alone report using the "PRIMERAWESOME-only" feature of POLYMSAT, which can also be used for non-microsatellite primers. This feature applies the *in silico* PCR routines of POLYMSAT on sequencing reads and existing primers specified by the user. In both cases, POLYMSAT when their list of predicted amplicons shares at least one NGS read. PRIMERAWESOME uses Clustal Omega (Sievers *et al.* 2011) to generate a multiple alignment for all reads in each primer group, to create a stack of aligned sequences.

Results and Discussion

EXPERIMENTAL DATA SET

High quality DNA was extracted from the blood of two individual Sumatran rhinoceros, both wild caught on the island of Sumatra and subsequently held in North American zoos. Experiments involving rhinoceros samples were conducted under IACUC approval number 15053. Each Sumatran rhino DNA sample was given a unique identifying barcoding tag before being sequenced simultaneously in one lane on the Illumina HiSeq V3 platform. A total of 30,556,224 sequencing reads were obtained, with

individual Dsu-33 producing 16,813,030 reads (average length of 410 bp) and individual Dsu-35 producing 13,743,194 reads (average length of 440 bp). Paired-end reads with overlapping sequence from each individual were merged using FLASH 1.2.8 to create FASTQ files containing 7,399,098 and 5,993,320 reads, for Dsu-33 and Dsu-35, respectively.

COMPUTATIONAL EFFICIENCY

Our goal for POLYMSAT and PRIMERAWESOME was to enable laboratories with modest computing budgets to leverage NGS data for microsatellite design. To ensure this goal was accomplished, we developed solely on a Windows laptop from 2008, with a 2.24GHz Intel Core 2 Duo processor. POLYMSAT designed polymorphic STR primers from Dsu-33 and Dsu-35 NGS reads in 6 hours on this laptop, using at most 100MB of memory during processing. To evaluate runtime on a modern computer, the POLYMSAT was run on a desktop with a 3.0GHz Intel Core i5, which finished within 1 hour.

POLYMSAT found 257,798 simple microsatellite arrays, containing at least 6 tandem copies of a 2-7 nucleotide motif, in 195,902 NGS reads. When simple motifs were within 50bp, they were combined into complex motifs. Of these, 64,148 repeat arrays without at least 40bp of non-repetitive flank were removed from consideration. An integrated version of Primer3 was used to design five primer sets per locus, with at least 18bp between the primer binding site and the repeat array, so that sequencing could uniquely identify each locus. Primers were successfully designed for 146,961 loci. Of these, 125,411 passed quality control screens that removed primers with multiple binding

sites in the genome and polymorphic primer binding sites. Of these, 64,594 contained sufficient sequencing coverage to resolve polymorphisms.

Since PolyMsat designs primers for each microsatellite-containing NGS read, multiple primers sets were designed for each genomic locus. PolyMsat performed *in-silico* PCR for all primers on all sequencing reads, and then grouped primer sets by putative amplicon sequence. PolyMsat then used BLAST to compare potential amplicon sequences for all primer pairs in each group. Primer sets that could amplify repetitive elements would likely have mismatches in their amplicons, so only primers with 95% similarity between all their (STR-masked) amplicons were retained. The resulting primers were predicted to amplify 19,921 unique loci.

In the Sumatran rhino, 3119 microsatellite loci were identified as polymorphic and 50 loci contained at least two alleles each supported by at least two reads (Table 2.1). This shows that ~2X coverage of two individuals is sufficient for polymorphic microsatellite identification but confidence in polymorphic markers would be greatly improved at higher coverage. Since PolyMsat does not rely on a genomic assembly to identify unique loci, overlapping microsatellite-containing regions are used to identify similar loci. Each microsatellite region requires a *ca.*100bp region (repeat array plus, per primer: 20bp for binding site + 20bp internal flank) to be present in a sequencing read to be identified. A large proportion of reads contains the microsatellite but lacks the contiguous *ca.* 100bp microsatellite region, so cannot be used to help genotype the microsatellite.

Of the primers designed, we chose 40 primer pairs for laboratory testing. Of the forty, 38 produced amplicons without excessive alleles (which is indicative of a repetitive

locus) in at least 2 DNA samples in an initial PCR. Of these, 35 of the markers submitted for fragment analysis produced genotype peak patterns that could be easily scored. At least 23 of the markers produced 2 or more alleles across the tested samples. While these markers and results validate the utility of the program, they are tangential to the development and description of the software, and will be described in a separate species-specific publication.

POLYMSAT improves on currently available microsatellite design software by enabling fast scanning of higher coverage next-generation sequencing data. Compared to existing software, such as Msatcommander (Faircloth 2008), Msatfinder (Thurston & Field 2005), RepeatMasker (Smit *et al.* 2013-2015) and SciRoKo (Kofler *et al.* 2007), PolyMsat is able to handle large-scale genomic sequences and leverage the high coverage to identify polymorphic loci within the sequenced individual(s). We successfully ran PolyMsat to design primers on ~2X coverage of the Sumatran rhino genome using a 2011 MacBook Air with a Core i5-2467M processor and 4GB of RAM in 6 hours. In addition, PolyMsat removes primers that bind disconcordantly, primers with polymorphic binding sites and microsatellite loci within repetitive elements. PolyMsat also provides an interactive report for the primers, which allows researchers to view each primer with its underlying sequencing reads, to make informed decisions before ultimately choosing primers.

Figures and Tables

Figure 2.1. Screenshot of a PRIMERAWESOME report.

The report is the output from POLYMSAT, which provides an interactive report browsing primers designed by POLYMSAT.

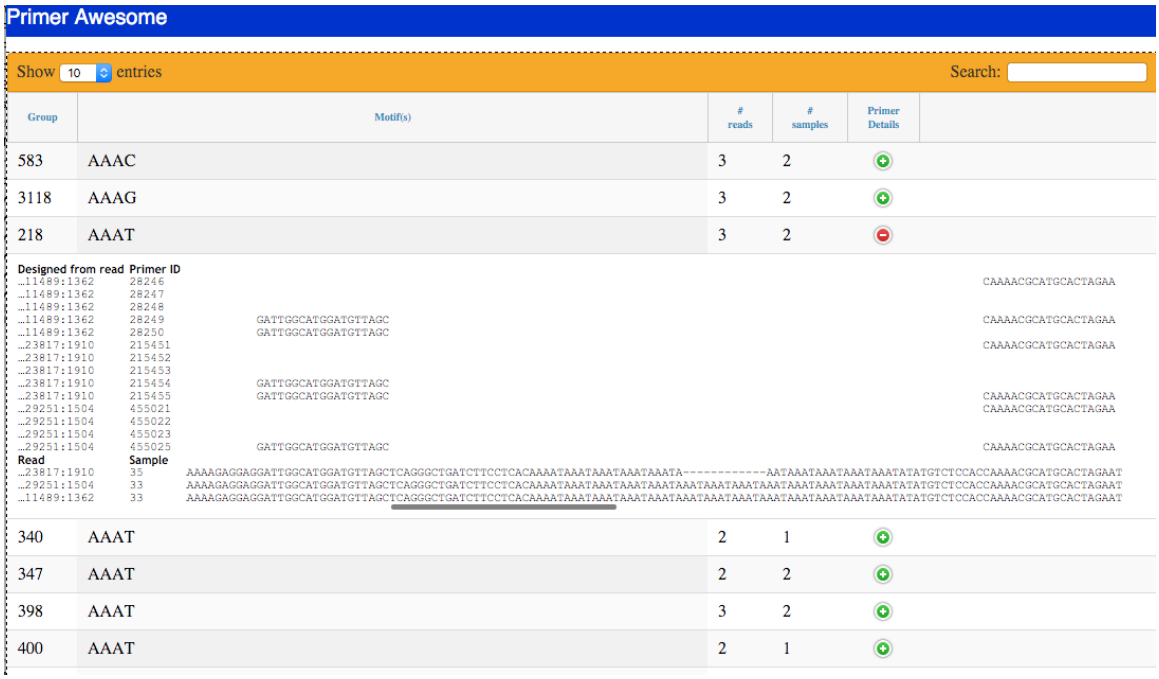

Figure 2.2. Screenshot of POLYMSAT.

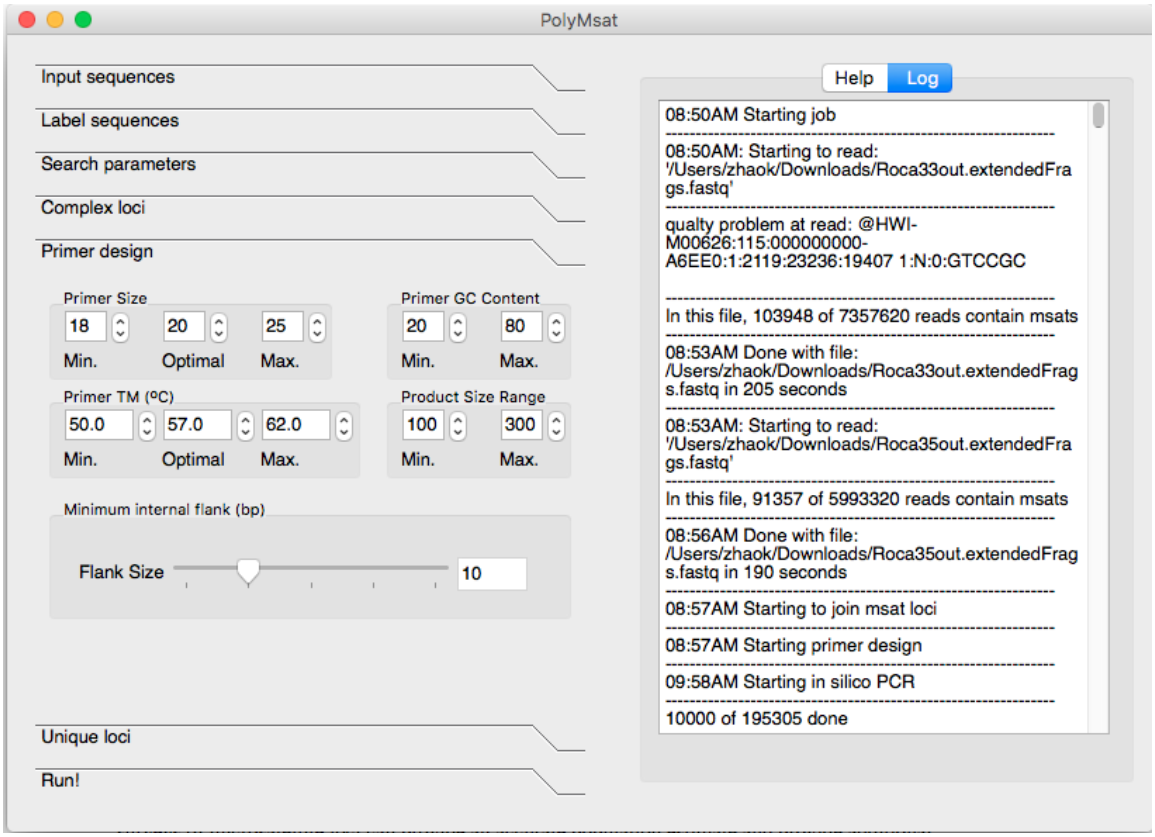

Figure 2.3. Flowchart for identifying polymorphic microsatellite loci and primer design.

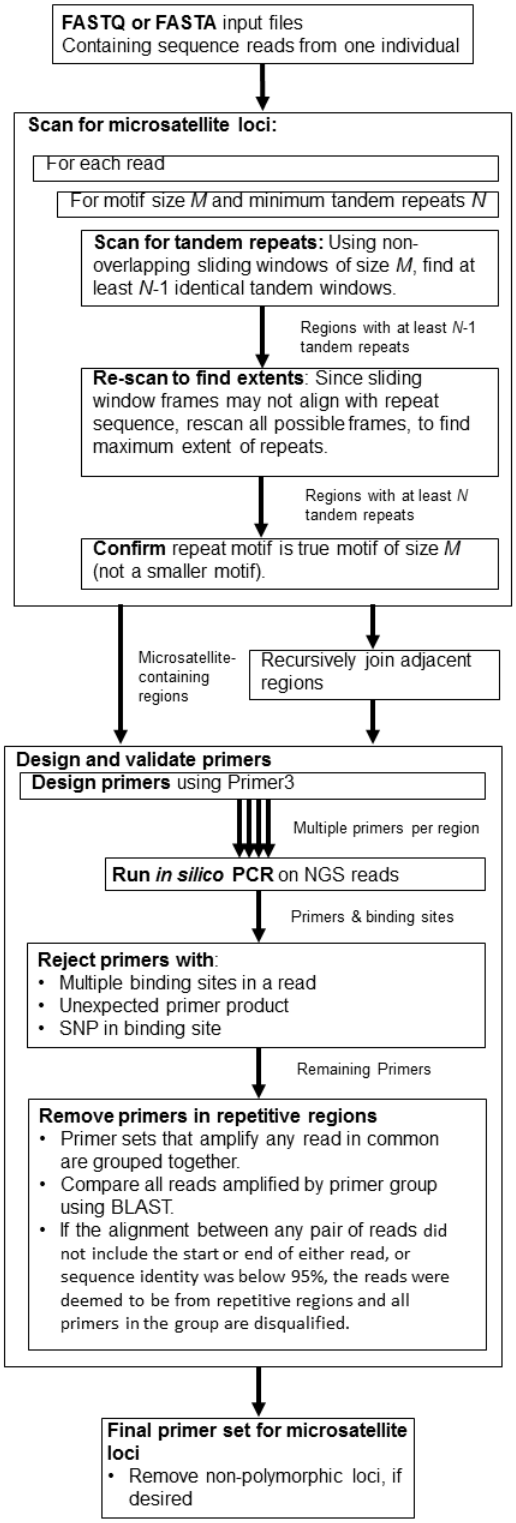

Table 2.1. Count of polymorphic loci identified by POLYMSAT. Polymorphic loci for microsatellite markers designed on next-generation sequencing reads from two Sumatran rhinoceros.

	Microsatellite Motif Size						
	2	**3**	**4**	**5**	**6**	**7**	**Complex motif**
Loci with ≥2 alleles	2667	83	327	6	0	0	36

CHAPTER 3: EVOLUTION OF MAMMALIAN LIMB DEVELOPMENT

Introduction

Vertebrate limb organogenesis requires intricate spatial-temporal modulation of protein synthesis through patterned dispersal of transcription and growth factors (Cooper & Sears). The homeobox gene family are characterized by their ~180 bp homeodomain, which encodes for a DNA binding protein domain, are often transcription factors (Holland *et al.* 2007). The Hox genes, a clustered family of homeobox genes involved in mammalian limb patterning, exemplify the surprising antiquity of this mechanism—Hox-like genes are found in organisms as basal as *Cnidaria* (Ferrier & Holland 2001). Mammalian genomes contain four homologous clusters of the *Hox* genes, which show strikingly high levels of concordance when compared to the *Hox* cluster in *Drosophila* (Duboule 2007; Ferrier & Holland 2001). Studies of traditional model organisms (chicken, mouse, pig) have shown that the regulatory mechanisms for limb development are relatively conserved. Thus, the morphological differences across tax are driven by changes in their enhancers and available transcription factors (TFs) (Stern 2000).

Changes in non-coding regulatory elements have been hypothesized to cause morphological differences that provide the material for evolutionary change (Stern 2000). In contrast to coding sequence changes, these mutations can modify the levels and the spatial-temporal patterns of gene expression, rather than changing gene function to affect all tissues (Cretekos *et al.* 2008). Notably, a single base pair mutation in the promoter *Shh* signaling gene leads to the development of an extra thumb/big toe in the mouse (preaxial polydactyly) (Maas & Fallon 2005).

In order to study the evolution of limb morphology genes, we examined the genomes of four mammalian species to study their divergent limb morphologies (mouse bat, pig and opossum). The mouse represents the pentadactyl mammalian condition with much known about its limb development process (Sears *et al.* 2011). Compared to other mammalian lineages, bats notably have a forelimb adapted for flight. Bats have elongated posterior digits (III-V) in their forelimb and retain the interdigital tissue to form the wing membrane (Cooper *et al.* 2012; Cretekos *et al.* 2008). Pigs possess an intermediate stage of digit reduction, a process experienced by half of all mammalian orders in their evolutionary history (Sears *et al.* 2011), making them an ideal candidate for studying limb evolution. Marsupials, including opossums, show rapid limb development and large forelimbs at birth (Richardson *et al.* 2009). They also serve as an outgroup to the three eutherian species. While these limb development differences are well characterized between mammalian lineages, the underlying regulatory changes are not well understood.

The first step of gene expression is DNA transcription by RNA polymerase II (RNAP). RNAP machinery is recruited by the promoter, a genetic region just upstream of the transcription start site (TSS) of a gene. Often promoters merely facilitate transcription at a relatively low level while distant regulatory regions called enhancers or *cis*-regulatory modules increase transcription by up to several orders of magnitude (Sinha *et al.* 2003). Transcription factors bind to enhancers then recruit co-activators and co-repressors to affect the overall transcription rate. These transcription factors are characterized by their binding site recognition sequences, known as motifs. Enhancers contain multiple binding sites for these transcription factors (Coffin 2004).

Enhancer functionality is modulated by the concentration of transcription factor proteins and accessibility of the enhancer, which is determined by histone modification. Regions marked by histone H3 acetylated at lysine 27 (H3k27ac) are highly correlated with active and accessible enhancer regions (Creyghton *et al.* 2010). Chromatin immunoprecipitation sequencing (CHIP-seq) has enabled genome-wide, high-resolution mapping of enhancer regions (Simmons *et al.* 2012). A recent study has mapped genomic regions marked by H3k27ac using CHIP-seq on mouse limb tissues during development (Cotney *et al.* 2013).

While these genome-wide scans identified locations for enhancers in limb tissue during development, they do not identify the transcription factors driving regulation nor do they identify the gene under regulation, since enhancers are often distant from the genes they regulate. These *cis*-regulatory modules (CRM) usually contain multiple binding sites for the same transcription factor and are highly conserved across species. These properties lend themselves to analyses involving the two commonly used computational methods for identifying transcription factor binding sites. Multiple alignments of orthologous genomic regions can be used to find over-represented and conserved patterns (He *et al.* 2009; Sinha 2007). Motif scanning algorithms use position weight matrixes (PWM) from known transcription factors to find strong putative binding sites within sliding windows of the genome. We used two motif-scanning programs implementing different algorithms: PatSer and Stubb. PatSer scans for informative matches to a PWM along a single genomic sequence and returns a log likelihood ratio (LLR) for a site interpreted to be the binding energy of the TF at that location (Duque 2013; Kim *et al.* 2010; Stormo *et al.* 1982). The presence of several strong binding sites

within a short (~500bp) span is interpreted as a CRM (Coffin 2004). We also used Stubb, which implements a Hidden Markov Model (HMM) and Expectation Maximization (EM) algorithm to determine an LLR for a sliding window that incorporates both strong and weak binding sites (Kim *et al.* 2010; Sinha *et al.* 2003).

Methods: Comparison of Expression Between Development Stages

Gene expression abundances from RNASeq experiments in mouse, bat, pig and opossum were acquired from our collaborators in the Sears Lab (Table 3.1) (GEO Accession: GSM1833591). RNA was extracted using the E.Z.N.A. Total RNA kit I and libraries RNASeq libraries were prepared with the Illumina TruSeq RNA Sample Preparation Kit. The libraries were sequenced on an Illumina HiSeq 2500 housed in the Roy G. Carver Biotechnology Center at the University of Illinois. The collaborators also used software to the Illumina adaptors and trimmed bases with a quality score below 20 at the 3' end of the read. For the bat, opossum and pig, they assembled the transcriptomes to their corresponding Ensembl reference genomes: GRCm38 (mouse), BROADO5 (opossum), and Sscrofa10.2 (pig). They generated a *de novo* assembly for the bat reads using Trinity (Grabherr *et al.* 2011).

We developed a pipeline using Python, which curated the RNASeq data, normalized metadata between samples and compared expression between organisms and across limb development stages. The data was stored in a Sqlite3 database to enforce formatting consistency and provide data access using Structured Query Language (SQL). The expression dataset included Ensembl gene identifiers for RNA transcripts for the mouse, opossum and pig. We used the Ensembl database to map homologous pig and bat

genes to the mouse genome to compare gene expression between species. The expression dataset provided for the bat genome included genes predicted by Tophat (Trapnell *et al.* 2009), which were aligned using translated nucleotide BLAST (tblastx) (Altschul *et al.* 1990) to find homologs in other species, including the mouse genome. The best-mapping mouse gene for each bat gene was assumed to be the homolog.

RNASeq quantifies expression in tissues by using NGS sequencing of reverse transcribed total RNA present in each sample. The sequencing depth of each nucleotide in the exome is roughly proportional to its frequency in the total RNA, such that longer genes have higher sequencing coverage than shorter genes (Love *et al.* 2014). Our collaborators normalized "raw counts" of genes mapped by the alignment software into FPKM values to allow for gene expression comparisons within samples.

We sought to characterize strong gene expression "direction changes" between stages using RNASeq within species. The fold change of FPKM values for each gene was calculated for each consecutive set of limb development stages for which data was available. Genes with low expression (FPKM < 1) in both stages were ignored, since these low-transcription genes are less likely to drive development. Expression fold changes between stages were binned into 20 equal-sized groups. The genes in the top and bottom bins were labeled as significantly up- and down-regulated, respectively.

The direction of expression for each mouse was compared to homologs in other species at each pair of development stages available. We developed a program in Python to identify genes with direction changes between stages that are found in one species but not in another; Direction changes between stages in the forelimb but not found in the hind limb (or vice versa). Selected examples were tested using whole mount *in situ*

hybridization to test the efficacy of this method.

Methods: Lineage Specific Transcription Factor Binding Site Changes

We sought to identify transcription factor binding site (herein: binding site) changes that contribute to the morphological differences that arise during mammalian limb development. Our strategy was to identify putative enhancer regions, across all available mammalian genomes, near to known limb development genes. These homologous sets of putative enhancers were scanned for transcription factor binding sites. Then a statistical test was applied to identify lineage-specific changes that may explain morphological differences in the limb.

Mouse Enhancer Homologs

Enhancers for limb development were acquired from two sources: known enhancers the literature and putative enhancers from chromatin marks. Known enhancers included the *HoxD* global control region A & B (Schneider *et al.* ; Spitz *et al.*), two highly conserved regions in a telomeric gene desert 385 Kb and 670 Kb away from the *HoxD* cluster (Andrey *et al.* 2013), the *prox* enhancer (Montavon *et al.* 2011), a conserved region upstream of *HoxA* (Lehoczky *et al.*) and an *shh* enhancer (Maas & Fallon 2005). The H3k27ac chromatin modification correlates with activated enhancers (Creyghton *et al.* 2010). A previous study used chromatin immunoprecipitation sequencing (ChIP-SEQ) to identify putatively activated enhancers in fetal mouse limb tissue at two developmental (Cotney *et al.* 2013). A separate study used ChIP-Seq to

identify Gli3 binding sites, a transcription factor associated with limb development, in fetal mouse limb tissue. We used a list of genes known to affect mouse limb development, which included all the genes in the four *Hox* clusters, *Fgf7*, *Wnt5a*, *Ext2*, *Evx1*, *Evx2*, and *Lnp*. We identified putatively activated enhancers (H3k27ac marks) and Gli3 binding sites within 100kb of the transcription start site for each of these genes (Cotney *et al.* 2013; Vokes *et al.* 2008).

We used the 60-species mouse conservation alignment from UCSC (Karolchik *et al.* 2014) to identify homologs to these putative and known mouse limb development genes. The alignment was created using pairwise Blastz alignments to the mouse mm10 reference assembly then combined into a multiple sequence alignment using multiz (Blanchette *et al.* 2004) (Karolchik *et al.* 2014). In effect, the alignment annotated homologous regions to the mouse genome. We developed an algorithm similar to the UCSC liftOver (Fujita *et al.* 2011) tool to allow discovery of enhancers that are both conserved and divergent. We used homologous sequence blocks as "seeds" or "anchors" for finding the homologous enhancers in other mammalian species. These anchors would enable searches for enhancers whose sequence has diverged through evolution or degraded through mutation.

The following method was applied to find both highly conserved and poorly conserved sequences between species, while removing spurious hits: For each mouse enhancer, we identified all multiple alignment blocks overlapping with its extents and processed each species contained in those blocks separately. Homologous enhancers that mapped to more than one multiple alignment block were checked to ensure they mapped concordantly across all blocks to the same genomic region. We extracted the sequence

from the enhancer homolog from the genomic sequence of the species using the multiple alignment blocks. The extracted sequence was ignored if the length was shorter than half the length of the mouse sequence, more than double the length of the mouse sequence or contained more than 20 missing base-calls.

TRANSCRIPTION FACTOR BINDING SITE PREDICTION

We determined the background transcription factor binding probability by resampling intronic genomic regions for each mammalian species. We sampled 2 kb genomic regions from each species using an uniform distribution across the genome. The region was not considered if the Ensembl annotation for the region contained a gene (Cunningham *et al.* 2015). The sampled regions were binned into four groups based on their GC content.

Motifs for transcription factors were collected from the JASPAR database (Portales-Casamar *et al.* ; Sandelin *et al.*) and from a recent experimental dataset (Weirauch *et al.* 2014). In addition we included the motif for ZBTB16**,** a transcription factor involved in limb development but not in either dataset. Motifs were converted into the Stormo format (Stormo *et al.* 1982).

We used Stubb (Sinha *et al.* 2003) to estimate the number of binding sites for each transcription factor putative limb enhancer homolog. We configured Stubb to scan overlapping 500 bp windows moved across each genomic region in 250 bp increments. I developed a Python tool to distribute tasks to and collect results from instances across ~600 CPUs on the University of Illinois Campus Cluster, using a MySQL database to coordinate jobs between the CPUs. The predicted number of binding sites for each TF in

each enhancer for each species is reported by Stubb. We compared each TF binding site count of each putative enhancer to the genomic window bin with a GC-content range containing that of the enhancer, to determine the empirical probability in the genomic background, which was subsequently converted into a z-score.

Brownian Motion Model for Enhancer Evolution

We used a Brownian motion model developed by Wei Yang in the Sinha laboratory to find enhancers with significant changes in the binding sites of a transcription factor in one lineage. The model envisions three scenarios: Spurious predicted transcription factor binding sites have high rates of evolution between lineages. True transcription factor binding sites would be evolutionarily constrained, thus the number of binding sites between species should be stable between lineages. Lineages with a functional change of transcription factor binding sites in an enhancer will have a sharp increase or decrease of binding sites, although the number of binding sites within the lineage will remain stable. Our model attempts to capture these three scenarios by calculating the evolution rate necessary for the binding site evolution at each edge of the phylogenetic tree, assuming that the rate was governed by a Brownian motion process.

The phylogenetic tree from the UCSC 60-way conservation track (Karolchik *et al.* 2014) was pruned to contain only the species with homologs for each enhancer. Using the number of binding sites predicted for a transcription factor and enhancer pair, we calculated the log likelihood of two models: The null model assuming a constant rate of binding site change in the phylogeny; and an alternative model assuming a constant rate of binding site change *except* at one edge of the phylogenetic tree. Lineage-specific

binding site changes should be captured by high log-likelihood ratios between these two models.

Transcription Factor Binding Site Viewer

I developed a transcription factor binding site viewer, using Python and JavaScript, to facilitate browsing of binding site changes between enhancers. The viewer was developed to improve upon aspects existing software, such as Insite (https://www.cs.utah.edu/~miriah/insite/), which only display predicted binding sites but not the underlying sequence change. For each transcription factor-enhancer pair to be displayed, I recomputed predicted binding sites using PatSer (Stormo *et al.* 1982). In contrast to Stubb, PatSer predicts the strength of binding sites across a genomic window, rather than an aggregate score accounting for both strong and weak binding sites. The discrete binding sites predicted, along with binding strength and location, are more conducive to graphical display. The results, along with the extracted multiple alignment for the region and associated metadata is exported in the JavaScript Object Notation (JSON) format. The viewer, which is written in JavaScript, using the JQuery & D3 frameworks and HTML provide an interactive browser for a mouse enhancer and its homologs across species (Figures 4.1a-b).

Results and Discussion

We sought to identify genes are differentially expressed between limb development stages in four species (mouse, pig, bat, opossum) using data from RNASeq experiments performed by the Sears laboratory (Table 3.1). We compared the expression

between developmental stages across the four species, when possible to find directional expression differences between species (Tables 3.2a-d). Genes with expression fold change in the top 5% between two development stages in one species and in the bottom 5% in another species are shown (Tables 3.2a-d). For example, *HOXA4* expression is significantly decreased between Wanek limb stages 2 & 3 in mouse and pig forelimb tissue while expression is significantly increased in opossum forelimb tissue (Table 3.2c). Our collaborators in the Sears lab validated the results of RNASeq experiments using whole mount *in situ* hybridization (WISH) for candidate genes selected based on prior knowledge. Their results, presented in Table 3.3, show that WISH corresponded well with the RNASeq analyses, although the genes selected for WISH experiments are based on prior knowledge and many of the genes did not reach the significance threshold for Tables 3.2a-d.

Our second goal for this experiment was to identify transcription factor changes between lineages that may drive the limb development differences between species. A few (2-5) high scoring candidates would be validated by collaborators *in situ.* We attempted to use a Brownian motion model to find transcription factor binding sites that evolve quickly in some lineages but are fixed in others. Unfortunately, we were not successful employing this strategy—an excessive number of binding sites were identified and even after adding empirical filters to remove false positives, we could not identify well qualified candidates for *in situ* experiments. We devised a simpler, *ad hoc* methodology: Calculate variance in the number of binding sites predicted by Stubb, for each transcription factor, in each enhancer homolog. Transcription factor-enhancer pairs with low variance across species (<0.05) but large changes (>0.70 binding sites gained or

lost relative to the mean) in the lineage containing a species of interest (mouse, bat, pig, opossum) would be considered (Table 3.4). We created a viewer (Figures 3.1a-b) to review the results from this method to consider transcription factor-enhancer pairs for further validation.

This study shows that RNASeq results correlate well to *in situ* hybridization for reporting transcription change between developmental stages. Our collaborators in the Sears laboratory validated nine *Hox* genes across several species and reported that the majority of the directional changes found by RNASeq aligned with their WISH results. The directional changes that could not be confirmed by WISH generally were transcribed at a lower quantity. The second phase of this study demonstrated the difficulty of finding lineage-specific changes in transcription factor binding sites, although we were able to find several candidates that are currently being validated by the Sears laboratory. We created improvements upon currently available transcription factor homology browsers available, which enabled more thorough review of candidate transcription factor binding sites.

Figures and Tables

Figure 3.1a. Screenshot of the transcription factor binding site viewer. A putative enhancer, based on an H3k27ac chromatin mark overlapping the *WNT5A* gene in mouse limb stage 3 tissue is shown. The black ticks on the lower left tracks show the predicted binding sites for SIX2 in several mammalian species (complete list of species not shown). The colored ticks, shown only for mouse, are predicted binding sites for all transcription factors for this enhancer. A multiple sequence alignment is shown (lower right) with a predicted binding site highlighted in yellow. The binding specificity at each nucleotide (motif) for the transcription factor is shown in the top right. Note the SNP between the mouse and rat sequence that affects the binding site.

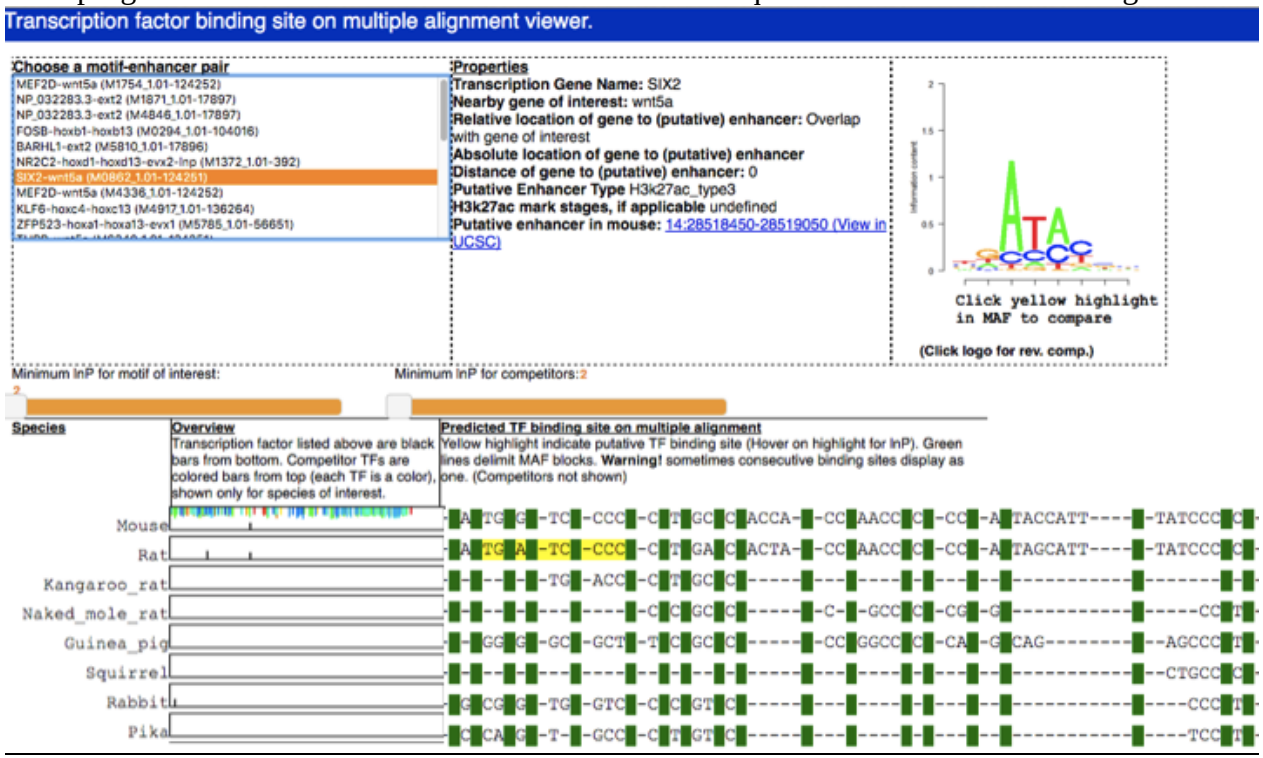

Figure 3.1b. Screenshot of the transcription factor binding site viewer. A putative enhancer, based on an H3k27ac chromatin mark overlapping the *WNT5A* gene in mouse limb stage 6 tissue is shown. The black ticks on the lower left tracks show the predicted binding sites for ZBTB12 in several mammalian species (complete list of species not shown). The colored ticks, shown only for mouse, are predicted binding sites for all transcription factors for this enhancer. The binding specificity at each nucleotide (motif) for the transcription factor is shown in the top right. A multiple sequence alignment is shown (lower right) with a predicted binding site highlighted in yellow.

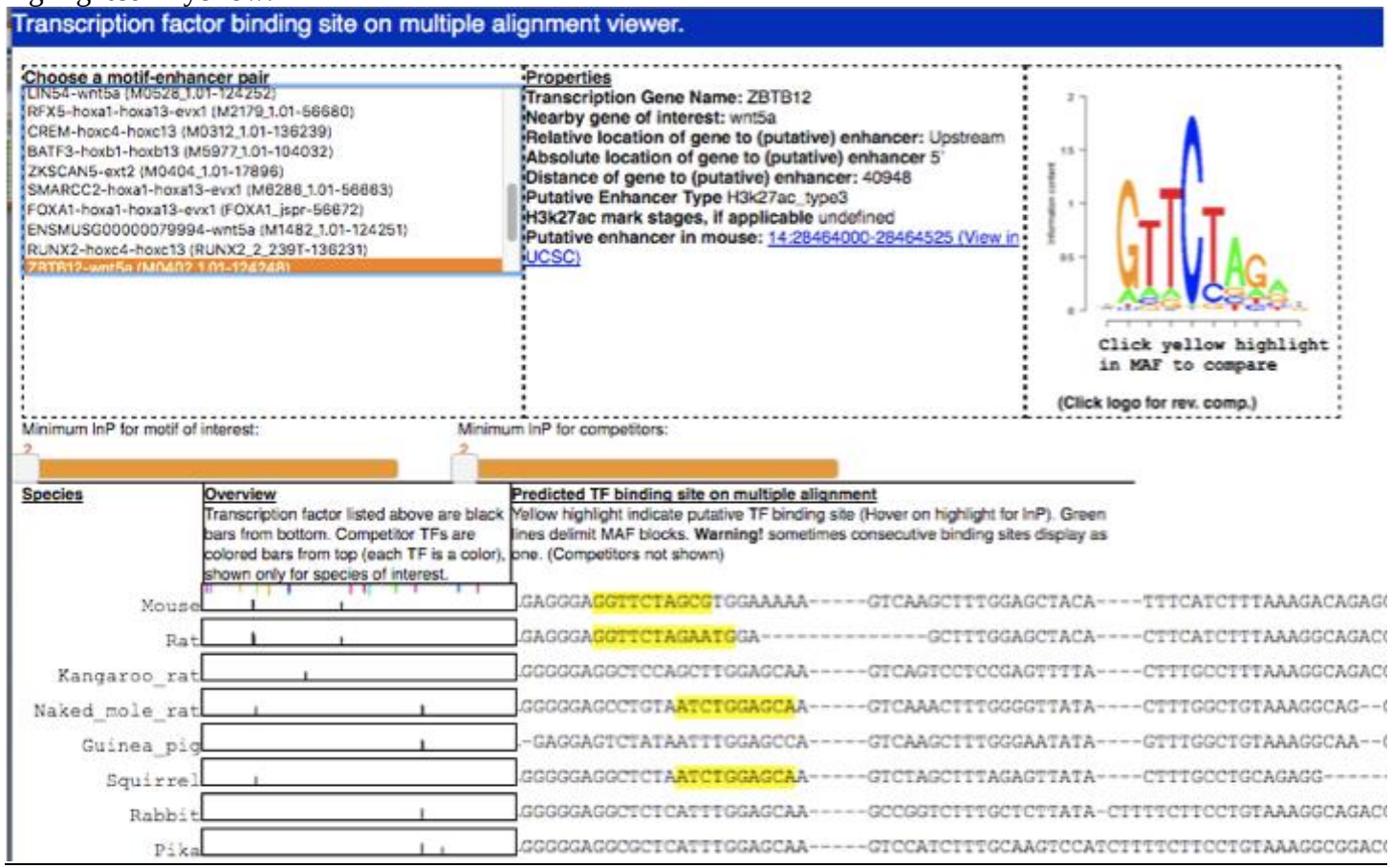

Table 3.1. **Limb samples used for RNASeq experiments,** indicating standard Wanek stages (Wanek *et al.*) (GEO Accession: GSM1833591). RNASeq results were available for both fore and hind limb (separately) for all samples except opossum forelimb at Wanek stage 2 (Opossum Stage 27).

Wanek Stage	**Mouse (*M. musculus*)**	**Pig (*S. scrofa)***	**Bat (*M. lucifugus)***	**Opossum (*M. domestica*)**
Stage 2	E10	n/a	Stage 13	Stage 30 (HL)
Stage 3/4	E11	E22	Stage 14	Stages 28 (FL)/31 (HL)
Stage 6	E12	E26	Stage 15	Stages 29 (FL)/31 (HL)

Table 3.2a. **Genes with directional differences in expression between Wanek stages 2 & 3 in the forelimb tissue.** Expression shown for Stage 2 ->Stage 3 as FPKM. Fold change between stage 2 & stage 3 is in parentheses. (Gene expression data not available for bat and opossum.) Blank cells indicate that expression for the homolog was not available.

Gene Name	Mouse	Pig	Bat	Opossum
ANO4	1.23->0.07 (0.1)		0.78->2.56 (3.3)	
GRIA3	5.35->0.43 (0.1)		0.23->1.58 (7.0);2.27->16.16 (7.1)	
MYH14	0.38->3.41 (9.0)		1.67->0.0 (0.0)	
MYL3	38.85->0.84 (0.0)		0.0->2.24 (inf)	
NTNG2	1.9->0.18 (0.1)		0.73->1.83 (2.5)	

Table 3.2b. **Genes with directional differences in expression between Wanek stages 2 & 3 in the hindlimb tissue.** Expression shown for Stage 2->Stage 3 as FPKM. Fold change between stage 2 & stage 3 is in parentheses. Blank cells indicate that expression for the homolog was not available.

Gene Name	Mouse	Pig	Bat	Opossum
ACOT6			1.2->6.54 (5.5)	1.15->0.6 (0.5)
ADH6B	1.26->0.33 (0.3)			1.92->16.95 (8.8)
ARL9	0.15->1.34 (9.2)			1.71->0.73 (0.4)
ATP6V1G3	0.04->1.86 (43.8)			2.17->0.86 (0.4)
CXCL5	1.05->0.44 (0.4)			0.12->1.21 (9.7)
DTX1	0.53->3.27 (6.2)		0.81->1.36 (1.7)	1.15->0.22 (0.2)
FAM183B	1.35->0.15 (0.1)			47.35->107.65 (2.3)
FAM46A	1.04->0.52 (0.5)			0.7->1.87 (2.7)
FOXF2	4.96->0.92 (0.2)			2.81->11.04 (3.9)
FXYD7	0.04->1.43 (35.3)			17.61->10.35 (0.6)
GABRP	0.47->4.35 (9.3)			8.36->4.75 (0.6)
GM22333	144.49->0.0 (0.0)			0.0->71.88 (inf)
GM22711	1101.6->488.57 (0.4)			0.0->34.8 (inf)
GM23969	0.0->524.46 (inf)			583.74->89.19 (0.2)
GM24336	0.0->28.1 (inf)			260.98->164.05 (0.6)
GM25291	60.69->12.06 (0.2)			57.73->314.17 (5.4)
GM25394	0.0->36.88 (inf)			87.36->24.63 (0.3)
GM25492	188.5->24.05 (0.1)			35.73->94.82 (2.7)
GM25579	112.57->53.73 (0.5)			0.0->145.02 (inf)
GM25776	184.05->85.21 (0.5)			0.0->43.78 (inf)
GM26079	0.0->54.89 (inf)			262.5->0.0 (0.0)
GM26351	0.0->5.0 (inf)			3.19->0.16 (0.1)
GPR22			2.1->5.22 (2.5)	1.51->0.65 (0.4)
HAPLN1	14.26->7.19 (0.5)			2.58->6.87 (2.7)
HBB-BH2	1.15->0.36 (0.3)			3.83->8.88 (2.3)
HSPB2			1.0->2.72 (2.7)	1.89->0.29 (0.2)
ISLR2	0.11->1.87 (16.6)			2.39->0.52 (0.2)
KLF15	3.02->1.4 (0.5)			0.25->1.09 (4.4)
KRT13	5.34->2.22 (0.4)			1.98->7.66 (3.9)
LGALS2	2.51->0.0 (0.0)			0.85->2.41 (2.8)
LOXL1	1.29->7.4 (5.7)			26.9->12.71 (0.5)

Table 3.2b (cont.) Genes with directional differences in expression between Wanek stages 2 & 3 in the hindlimb tissue. Expression shown for Stage 1->Stage 2 as FPKM. Fold change between stage 2 & stage 3 is in parentheses. Blank cells indicate that expression for the homolog was not available.

Gene Name	Mouse	Pig	Bat	Opossum
NKAIN4	1.75->7.54 (4.3)			3.39->1.4 (0.4)
PAX1	0.89->4.57 (5.1)		4.72->0.56 (0.1) 4.72->0.56 (0.1)	2.81->0.24 (0.1)
PLXNA4	0.48->2.25 (4.7)			1.92->0.89 (0.5)
REL			2.14->0.0 (0.0)	0.47->1.31 (2.8)
RGS16	0.57->3.94 (6.9)			1.1->0.09 (0.1)
RRAD	0.46->2.02 (4.4)			2.15->0.7 (0.3)
RSPO1	11.55->5.5 (0.5)		0.0->1.62 (inf)	10.63->5.01 (0.5)
S100A8	1.34->0.11 (0.1)			0.0->2.58 (inf)
SCARNA3A	0.0->39.24 (inf)			66.68->33.34 (0.5)
TCFL5	1.61->0.62 (0.4)			0.38->1.28 (3.4)
TSPAN11	2.18->15.39 (7.1)			1.26->0.69 (0.5)
TTR	7.9->0.1 (0.0)			0.19->1.23 (6.5)
UNC45B	0.13->1.23 (9.7)			1.37->0.59 (0.4)

Table 3.2c. Genes with directional differences in expression between Wanek stages 3 & 6 in the forelimb tissue. Expression shown for Stage 3->Stage 6 as FPKM. Fold change between stage 3 & stage 6 is in parentheses. Blank cells indicate that expression for the homolog was not available.

Gene Name	Mouse	Pig	Bat	Opossum
5430435G22RIK		0.63->1.69 (2.7)		1.53->0.7 (0.5)
ACSM3	0.14->20.56 (150.5)			1.05->0.5 (0.5)
ADAMTSL5	2.39->0.34 (0.1)	0.18->1.12 (6.3)		
ADCYAP1		21.84->6.31 (0.3)		0.18->2.58 (14.4)
AFP		682.29->1.5 (0.0)		0.12->3.15 (27.3)
AMBP		57.37->0.26 (0.0)		0.01->4.84 (732.9)
ANG	1.14->0.51 (0.4)	17.85->4.42 (0.2)		0.36->1.86 (5.1)
ANG2		17.85->4.42 (0.2)		0.36->1.86 (5.1)
ANG4		17.85->4.42 (0.2)		0.36->1.86 (5.1)
ANG5		17.85->4.42 (0.2)		0.36->1.86 (5.1)
ANG6		17.85->4.42 (0.2)		0.36->1.86 (5.1)
ARG1	1.45->1.26 (0.9)	1.49->0.02 (0.0)		0.21->1.88 (9.2)
ASGR1		9.05->0.0 (0.0)		0.36->1.47 (4.0)
ATF3	0.92->1.07 (1.2)	1.13->10.53 (9.3)		3.43->1.26 (0.4)
BGLAP	3.48->0.12 (0.0)	6.86->5.23 (0.8)		1.67->7.61 (4.6)
CHRNB4		2.06->0.84 (0.4)		0.05->1.47 (31.4)
CLEC18A		3.27->0.13 (0.0)		0.13->2.03 (15.1)
CLIC6		1.0->0.79 (0.8)	0.24->3.57 (14.9)	1.65->0.06 (0.0)
CLVS1	1.26->0.0 (0.0)			0.15->2.76 (18.8)
CXCL10	0.12->1.59 (13.0)			4.41->0.48 (0.1)
DAPL1	1.09->0.85 (0.8)	0.76->2.5 (3.3)		3.93->1.7 (0.4)
E030030I06RIK	1.38->19.23 (13.9)			6.13->3.0 (0.5)
EGFLAM	4.52->10.02 (2.2)	8.17->3.8 (0.5)		1.35->4.14 (3.1)
FABP1		16.57->0.62 (0.0)		0.67->4.27 (6.4)
FGA		18.55->0.0 (0.0)		0.07->4.13 (59.9)
FGB		3.62->0.03 (0.0)		0.2->2.8 (14.3)
FGG		1.05->0.0 (0.0)		0.1->3.58 (35.2)
FOXD1	2.95->0.19 (0.1)			0.47->4.53 (9.6)
FSCN3			2.24->0.0 (0.0) 2.07->0.0 (0.0)	0.25->1.21 (4.8)
FXYD7	4.62->1.16 (0.3)			8.6->72.4 (8.4)
GAL	2.58->0.28 (0.1)			0.35->3.37 (9.7)
GAP43	46.77->21.63 (0.5)	23.81->9.23 (0.4)		0.98->17.48 (17.7)
GM20091	17.1->0.0 (0.0)	46.09->115.66 (2.5)		10065.48->9490.47 (0.9)

Table 3.2c (cont.) Genes with directional differences in expression between Wanek stages 3 & 6 in the forelimb tissue. Expression shown for Stage 3->Stage 6 as FPKM. Fold change between stage 3 & stage 6 is in parentheses. Blank cells indicate that expression for the homolog was not available.

Gene Name	Mouse	Pig	Bat	Opossum
GM22154	0.0->42.02 (inf)	161.15->0.0 (0.0)		149.64->210.72 (1.4)
GM22357		0.0->382.92 (inf)		287.82->0.0 (0.0)
GM22442	0.0->68.13 (inf)	62.52->0.0 (0.0)		87.85->150.22 (1.7)
GM22661	0.0->58.97 (inf)	76.15->0.0 (0.0)		78.16->26.29 (0.3) 155.41->174.9 (1.1)
GM22680		0.0->1421.81 (inf)		1166.88->0.0 (0.0)
GM22684		80.38->0.0 (0.0)		46.77->211.31 (4.5)
GM22797		80.38->0.0 (0.0)		46.77->211.31 (4.5)
GM23297	0.0->48.6 (inf)	93.83->195.97 (2.1)		104.01->0.0 (0.0)
GM23318	0.0->91.87 (inf)	165.22->0.0 (0.0)		
GM23679		101.01->423.67 (4.2)		1157.0->516.4 (0.4)
GM23734		0.0->151.7 (inf)		218.95->0.0 (0.0)
GM23772		101.01->423.67 (4.2)		1157.0->516.4 (0.4)
GM23925		80.38->0.0 (0.0)		46.77->211.31 (4.5)
GM23946	0.0->37.33 (inf)			18.34->5.04 (0.3)
GM24031	0.0->36.83 (inf)	80.38->0.0 (0.0)		46.77->211.31 (4.5)
GM24163		80.38->0.0 (0.0)		46.77->211.31 (4.5)
GM24201	79.86->0.0 (0.0)	101.01->423.67 (4.2)		1157.0->516.4 (0.4)
GM24407	0.0->96.55 (inf)			100.63->23.4 (0.2)
GM24411	101.97->128.7 (1.3)	80.38->0.0 (0.0)		46.77->211.31 (4.5)
GM24494	0.0->268.53 (inf)			694.51->338.33 (0.5)
GM24507	0.0->46.01 (inf)			91.18->0.0 (0.0)
GM24601	0.0->27.89 (inf)	95341.0->1405250.0 (14.7)		32.04->0.0 (0.0)
GM24613	0.0->72.84 (inf)			306.24->123.79 (0.4)
GM24924	34.11->0.0 (0.0)	0.0->64.56 (inf)		
GM25187		80.38->0.0 (0.0)		46.77->211.31 (4.5)
GM25283		0.0->942.22 (inf)		104.5->0.0 (0.0)

Table 3.2c (cont.) Genes with directional differences in expression between Wanek stages 3 & 6 in the forelimb tissue. Expression shown for Stage 3->Stage 6 as FPKM. Fold change between stage 3 & stage 6 is in parentheses. Blank cells indicate that expression for the homolog was not available.

Gene Name	Mouse	Pig	Bat	Opossum
GM25402		80.38->0.0 (0.0)		46.77->211.31 (4.5)
GM25428		101.01->423.67 (4.2)		1157.0->516.4 (0.4)
GM25492	164.75->42.25 (0.3)	74.76->145.54 (1.9)		8.14->169.15 (20.8)
GM25592		80.38->0.0 (0.0)		46.77->211.31 (4.5)
GM25683		80.38->0.0 (0.0)		46.77->211.31 (4.5)
GM25776	0.0->65.5 (inf)			57.68->0.0 (0.0)
GM25782	2.56->0.0 (0.0)			37.32->121.13 (3.2)
GM25848	0.0->65.07 (inf)			57.68->0.0 (0.0)
GM25966	0.0->202.66 (inf)	460.24->129.7 (0.3)		33.0->189.89 (5.8)
GM26079	0.0->39.7 (inf)			76.61->27.19 (0.4)
GM26158	0.0->74.39 (inf)			153.19->0.0 (0.0)
GM26244	34.11->0.0 (0.0)	0.0->64.56 (inf)		
GM26440	0.0->39.63 (inf)			306.24->123.79 (0.4)
GNG13	8.89->0.98 (0.1)	0.0->23.45 (inf)		3.71->9.94 (2.7)
GSX2			4.47->0.0 (0.0) 4.47->0.0 (0.0)	0.07->1.11 (15.5)
HIST1H3A	5.77->1.54 (0.3)	3.79->13.69 (3.6)		
HIST1H3G	3.47->0.84 (0.2)	0.0->2.85 (inf)		
HMX1	3.57->0.04 (0.0)			0.23->2.62 (11.4)
HOXA4	5.72->1.51 (0.3)	2.19->1.41 (0.6)		2.47->10.77 (4.4)
HOXB3	0.59->2.43 (4.1)	14.6->0.75 (0.1)		2.8->14.41 (5.1)
HOXB4	1.58->0.49 (0.3)	7.7->0.44 (0.1)		2.1->7.47 (3.6)
HOXB5	3.33->0.58 (0.2)	3.73->0.1 (0.0)		2.8->16.59 (5.9)
HOXB6	6.58->0.72 (0.1)	1.94->0.41 (0.2)		1.34->6.8 (5.1)
HOXB7	7.43->1.55 (0.2)	2.07->0.88 (0.4)		1.33->4.65 (3.5)
HOXC4	9.62->3.0 (0.3)	14.34->1.74 (0.1)		3.91->13.7 (3.5)
HS3ST2		1.03->0.41 (0.4)		0.16->1.38 (8.4)
IGFBP1		2.25->0.0 (0.0)		0.17->1.2 (6.9)
IHH		0.2->1.32 (6.6)	0.0->43.77 (inf)	2.3->0.06 (0.0)
IKZF1	0.24->6.46 (26.4)			2.14->0.97 (0.5)
IL17RE		3.59->9.64 (2.7)		1.7->0.64 (0.4)
ITIH2		13.06->0.07 (0.0) 13.19->0.2 (0.0)		0.04->4.89 (128.1)
KCP	33.99->17.92 (0.5)	11.78->4.61 (0.4)	6.72->4.88 (0.7) 6.72->4.88 (0.7)	1.62->5.04 (3.1)

Table 3.2c (cont.) Genes with directional differences in expression between Wanek stages 3 & 6 in the forelimb tissue. Expression shown for Stage 3->Stage 6 as FPKM. Fold change between stage 3 & stage 6 is in parentheses. Blank cells indicate that expression for the homolog was not available.

Gene Name	Mouse	Pig	Bat	Opossum
KLK7			0.0->2.01 (inf)	5.0->0.94 (0.2)
LRP2			0.0->2.51 (inf)	1.48->0.34 (0.2)
MIR1839	71.22->0.0 (0.0)	0.0->62.92 (inf)		
MIR199A-1		527.09->0.0 (0.0)		67.18->274.16 (4.1)
MYL2	2.68->0.55 (0.2)	2.19->3.86 (1.8)		0.08->56.21 (747.4)
MYL7	2.75->0.13 (0.0)	8.88->9.97 (1.1)		1.1->91.17 (82.6)
NACAD	5.56->1.35 (0.2)			0.08->1.36 (16.9)
NCMAP	0.16->1.91 (12.3)	3.29->35.63 (10.8)		1.32->0.46 (0.3)
NEFL	1.2->0.27 (0.2)			0.73->15.31 (20.9)
NPFF	8.13->1.15 (0.1)	1.76->7.15 (4.1)		
NPM2	1.95->0.45 (0.2)	0.29->3.53 (12.2)		0.27->1.22 (4.4)
PAX1	22.05->6.34 (0.3)	3.23->0.36 (0.1)	7.28->2.0 (0.3) 7.28->2.0 (0.3)	2.09->17.69 (8.5)
PLG		21.19->0.02 (0.0)		0.1->1.09 (11.3)
PRSS53		0.43->1.14 (2.6)		1.0->0.38 (0.4)
RAB6B		2.78->1.11 (0.4)	3.9->5.2 (1.3)	0.32->1.98 (6.2)
RANBP3L	0.0->1.05 (inf)			1.08->0.12 (0.1)
RMRP	121.11->95.11 (0.8)	0.0->21.24 (inf)		36.0->15.53 (0.4)
RNY1	541.16->75.16 (0.1)	148.74->1467.15 (9.9)		
RNY3	564.69->98.17 (0.2)	0.0->1264.3 (inf)		101.25->0.0 (0.0)
S100A8	0.0->2.1 (inf)	1.57->0.44 (0.3)		
SERPINC1		10.45->0.38 (0.0)		0.21->1.92 (9.3)
SLC35D3		3.06->0.6 (0.2)		0.36->1.6 (4.4)
SNCG	10.48->1.51 (0.1)			0.2->4.64 (22.8)
SNORA30	0.0->96.73 (inf)			88.1->0.0 (0.0)
SNORA36B	0.0->35.54 (inf)	224.0->70.53 (0.3)		107.26->68.51 (0.6)
SNORA75	230.85->14.53 (0.1)			233.96->1168.38 (5.0)
SNORA78	313.65->0.0 (0.0)			0.0->69.6 (inf)
SNORD15A	0.0->33.84 (inf)			45.76->16.11 (0.4)
SNORD66	0.0->1307.77 (inf)	637.9->0.0 (0.0)		
SNORD8	0.0->184.17 (inf)			218.95->0.0 (0.0)
SSTR1	0.04->1.26 (33.6)	0.25->3.71 (14.8)		2.1->0.33 (0.2)
STMN3	1.08->0.13 (0.1)			1.39->25.29 (18.2)

Table 3.2c (cont.) Genes with directional differences in expression between Wanek stages 3 & 6 in the forelimb tissue. Expression shown for Stage 3->Stage 6 as FPKM. Fold change between stage 3 & stage 6 is in parentheses. Blank cells indicate that expression for the homolog was not available.

Gene Name	Mouse	Pig	Bat	Opossum
STMN4	1.26->0.35 (0.3)			0.12->3.71 (30.2)
STYK1	0.07->1.24 (16.8)	1.76->1.12 (0.6)		1.35->0.31 (0.2)
TCAP	2.67->0.27 (0.1)	0.77->1.85 (2.4)		3.45->7.66 (2.2)
TGFA	0.14->1.18 (8.5)	2.54->5.99 (2.4)		11.22->4.37 (0.4)
TMEFF2		8.17->0.9 (0.1)		0.19->1.23 (6.5)
TMEM154	0.04->1.58 (38.6)	3.95->4.71 (1.2)		1.21->0.45 (0.4)
TNFAIP8L3		1.97->0.0 (0.0)		0.48->2.19 (4.6)
VGLL3	1.19->1.96 (1.7)	2.08->6.76 (3.2)		4.83->2.0 (0.4)
WNT2	1.75->0.47 (0.3)			0.27->1.97 (7.3)
WT1	1.51->0.26 (0.2)	2.74->0.15 (0.1)		1.21->4.78 (4.0)

Table 3.2d. Genes with directional differences in expression between Wanek stages 3 & 6 in the hindlimb tissue. Expression shown for Stage 3->Stage 6 as FPKM. Fold change between stage 3 & stage 6 is in parentheses. Blank cells indicate that expression for the homolog was not available.

Gene Name	Gene Name	Gene Name	Gene Name	Gene Name
4930546H06 RIK	127.49->54.82 (0.4)	0.46->2.15 (4.7)		79.51->164.56 (2.1)
ACTA1		5.81->1.18 (0.2)		0.92->4.59 (5.0)
ACTG2		1.44->0.23 (0.2)		0.29->1.16 (4.0)
ALDH1A1		50.49->1.99 (0.0)		2.55->8.29 (3.3)
ALDH1A2	17.87->14.73 (0.8)	27.33->13.15 (0.5)		2.71->9.94 (3.7)
ALDH1A7		50.49->1.99 (0.0)		2.55->8.29 (3.3)
ATP6V1G3	1.41->2.93 (2.1)	0.93->2.65 (2.9)		3.5->0.34 (0.1)
C1QC	1.92->3.3 (1.7)	1.22->0.63 (0.5)		4.4->10.91 (2.5)
C1QTNF3	1.62->1.61 (1.0)	15.67->3.54 (0.2)		0.67->5.16 (7.7)
C1QTNF5		2.52->1.07 (0.4)		1.31->7.22 (5.5)
CDH10	1.58->0.7 (0.4)	0.75->1.78 (2.4)		
CDNF	0.27->1.75 (6.5)	1.37->0.0 (0.0)		
COL11A2	5.57->2.39 (0.4)	1.38->18.62 (13.5)		0.52->1.66 (3.2)
COL8A2	2.66->1.19 (0.4)	1.28->7.95 (6.2)		3.72->9.53 (2.6)
COLGALT2	1.31->0.65 (0.5)	0.82->2.6 (3.2)		0.37->1.06 (2.9)
CSMD3		0.01->1.65 (146.6)	1.88->0.0 (0.0)	
CXCL10	0.49->1.29 (2.6)			1.83->0.75 (0.4)
CYP1B1	1.38->0.97 (0.7)	1.45->0.51 (0.3)		0.28->1.08 (3.8)
D330045A20 RIK	0.34->1.91 (5.6)	2.3->0.88 (0.4) 1.97->0.89 (0.5)		
EMCN	19.24->17.59 (0.9)	42.36->21.33 (0.5)		1.28->3.0 (2.4)
ERICH2	2.06->0.14 (0.1)	75.62->201.4 (2.7)		
ERMAP	7.4->7.35 (1.0)		2.67->0.42 (0.2)	0.57->1.45 (2.5)
ETV2	1.09->0.29 (0.3)	1.02->2.89 (2.8)		
FBP1		1.37->0.49 (0.4)		0.23->1.1 (4.7)
FXYD2	0.64->1.53 (2.4)	158.96->21.96 (0.1)		0.0->41.51 (inf)
FYB	0.64->2.97 (4.7)		2.54->0.0 (0.0)	
GJB6	7.63->2.73 (0.4)			0.11->2.16 (19.0)
GLIS3	3.01->1.42 (0.5)		0.14->2.16 (15.4)	
GM10037	0.0->1.2 (inf)	13.35->4.4 (0.3) 22.15->8.02 (0.4) 31.96->17.08 (0.5)		2.8->1.68 (0.6)
GM21983	1.26->0.36 (0.3)	4.45->79.2 (17.8)		
GM22156		0.0->80.47 (inf)		53.08->0.0 (0.0)

Table 3.2d (cont.) Genes with directional differences in expression between Wanek stages 3 & 6 in the hindlimb tissue. Expression shown for Stage 3->Stage 6 as FPKM. Fold change between stage 3 & stage 6 is in parentheses. Blank cells indicate that expression for the homolog was not available.

Gene Name	Mouse	Pig	Bat	Opossum
GM22247	0.0->102.4 (inf)			86.49->26.55 (0.3)
GM22684		0.0->68.97 (inf)		103.94->36.59 (0.4)
GM22711	488.57->237.47 (0.5)			14.91->99.68 (6.7)
GM22786	0.0->123.81 (inf)			208.9->0.0 (0.0)
GM22797		0.0->68.97 (inf)		103.94->36.59 (0.4)
GM22806	197.2->62.73 (0.3)			0.0->62.92 (inf)
GM23202	284.69->174.95 (0.6)	0.0->257.66 (inf)		887.05->226.15 (0.3)
GM23262	85.2->130.96 (1.5)	0.0->501.0 (inf)		351.38->112.7 (0.3)
GM23297	191.56->66.0 (0.3)	0.0->255.82 (inf)		
GM23723	0.0->161.1 (inf)	245.67->58.8 (0.2)		206.2->45.14 (0.2)
GM23758		139.03->343.72 (2.5)		130.49->46.04 (0.4)
GM23893		0.0->65.08 (inf)		103.94->36.59 (0.4)
GM23925	133.37->0.0 (0.0)	0.0->68.97 (inf)		103.94->36.59 (0.4)
GM23971		102.11->318.49 (3.1)		15.45->0.0 (0.0)
GM24031		0.0->68.97 (inf)		103.94->36.59 (0.4)
GM24044	348.07->79.22 (0.2)	0.0->33.89 (inf)		421.1->280.21 (0.7)
GM24163		0.0->68.97 (inf)		103.94->36.59 (0.4)
GM24411	162.97->111.52 (0.7)	0.0->68.97 (inf)		103.94->36.59 (0.4)
GM24613	0.0->115.94 (inf)			243.99->46.52 (0.2)
GM25187		0.0->68.97 (inf)		103.94->36.59 (0.4)
GM25402		0.0->68.97 (inf)		103.94->36.59 (0.4)
GM25492	24.05->104.57 (4.3)	129.34->62.81 (0.5)		102.72->0.0 (0.0)
GM25541	150.67->72.74 (0.5)	0.0->64.17 (inf)		30.6->15.17 (0.5)
GM25592		0.0->68.97 (inf)		103.94->36.59 (0.4)
GM25683		0.0->68.97 (inf)		103.94->36.59 (0.4)

Table 3.2d (cont.) Genes with directional differences in expression between Wanek stages 3 & 6 in the hindlimb tissue. Expression shown for Stage 3->Stage 6 as FPKM. Fold change between stage 3 & stage 6 is in parentheses. Blank cells indicate that expression for the homolog was not available.

Gene Name	Mouse	Pig	Bat	Opossum
GM25820	0.0->64.95 (inf)			41.9->0.0 (0.0)
GM25945	37.76->0.0 (0.0)	0.0->61.63 (inf)		163.89->97.43 (0.6)
GM25966	214.86->0.0 (0.0)	0.0->109.65 (inf)		84.03->70.01 (0.8)
GM26003	0.0->61.71 (9.59057614405e+59)	151.1->0.0 (0.0)		103.94->36.59 (0.4)
GM26130	60.61->177.54 (2.9)	332.5->55.16 (0.2)		181.08->86.49 (0.5)
GM26265	66.94->0.0 (0.0)	0.0->221.84 (inf)		
GPHA2	0.36->1.17 (3.2)	1.92->0.89 (0.5)		
GPR64		0.34->1.12 (3.3)		1.09->0.11 (0.1)
GREM2	1.01->0.2 (0.2)			0.18->2.16 (11.8)
GRIA2	1.6->0.4 (0.3)	1.61->3.4 (2.1)		0.24->1.16 (4.8)
HIST1H4K	1.62->6.64 (4.1)	20.84->11.12 (0.5)		
HPN	3.06->2.46 (0.8)	3.36->1.4 (0.4)		0.6->1.4 (2.3)
HR	1.16->0.46 (0.4)			0.51->2.03 (4.0)
KCNQ5		3.64->8.62 (2.4)		2.25->0.85 (0.4)
KLHL14		1.23->0.32 (0.3)		0.08->1.15 (13.6)
LDHD		1.91->0.98 (0.5)		0.73->1.79 (2.5)
MAGIX		0.43->2.17 (5.1)		8.21->3.06 (0.4)
MAP2	20.12->8.83 (0.4)	3.11->5.84 (1.9)		0.46->1.94 (4.2)
MIR1949	90.74->48.15 (0.5)	0.0->52.81 (inf)		253.11->72.17 (0.3)
MYL3			1.01->0.0 (0.0)	0.68->2.82 (4.1)
MYOF	1.25->0.4 (0.3)	2.75->4.33 (1.6) 1.07->2.0 (1.9)	0.86->4.53 (5.2)	2.01->2.13 (1.1)
N-R5S88	0.0->107.83 (inf)	184.43->0.0 (0.0)		
NCCRP1		1.42->0.72 (0.5)		5.33->13.06 (2.5)
NCMAP		2.64->24.12 (9.1)		3.56->0.0 (0.0)
PAQR6	1.03->0.2 (0.2)	0.66->1.75 (2.7)		1.13->2.68 (2.4)
PAX1	4.57->1.88 (0.4)	1.39->1.19 (0.9)		0.24->1.54 (6.4)
PDZD7	1.5->0.77 (0.5)	0.25->1.09 (4.4)		0.22->1.08 (5.0)
PLA2G4E		1.85->0.44 (0.2)		0.13->3.08 (24.1)
PNCK	4.56->6.78 (1.5)	11.92->3.55 (0.3)		1.07->2.86 (2.7)
POSTN	9.01->2.42 (0.3)	10.88->25.59 (2.4)		88.29->51.72 (0.6)
PRELP	1.04->0.53 (0.5)	0.19->2.68 (14.3)		
PRRT2	3.0->0.87 (0.3)	2.13->5.59 (2.6)		
RUNX2	2.97->1.08 (0.4)	1.16->3.87 (3.3)		2.31->2.73 (1.2)

Table 3.2d (cont.) Genes with directional differences in expression between Wanek stages 3 & 6 in the hindlimb tissue. Expression shown for Stage 3->Stage 6 as FPKM. Fold change between stage 3 & stage 6 is in parentheses. Blank cells indicate that expression for the homolog was not available.

Gene Name	Mouse	Pig	Bat	Opossum
		1.36->7.31 (5.4)		
RXRG	1.01->0.51 (0.5)	0.83->2.43 (2.9)		
SCARNA13	10.0->2.19 (0.2)	7.89->36.05 (4.6)		6.05->7.38 (1.2)
SCARNA17		0.0->242.47 (inf)		404.38->167.79 (0.4)
SLC35G1	2.61->2.34 (0.9)	2.73->1.28 (0.5)		0.41->1.2 (2.9)
SNORA20	0.0->99.97 (inf)	449.91->362.85 (0.8)		511.05->165.79 (0.3)
SNORA36B	0.0->81.1 (inf)	358.44->0.0 (0.0)		86.49->26.55 (0.3)
SNORD100	0.0->544.42 (inf)	0.0->392.1 (inf)		370.67->0.0 (0.0)
SNORD104	6096.69->5786.76 (0.9)	1496.84->3708.92 (2.5)		1388.33->0.0 (0.0)
SNORD35B	0.0->442.55 (inf)			210.11->0.0 (0.0)
SRL	48.07->25.15 (0.5)	0.66->1.17 (1.8)		0.15->1.3 (8.6)
TBX22	1.32->0.64 (0.5)	0.34->6.9 (20.3)		
TCEAL7	1.23->0.52 (0.4)	23.82->60.95 (2.6)		
TEX12		1.49->0.04 (0.0)	0.39->2.28 (5.9)	
TH	0.43->1.08 (2.5)	3.3->0.2 (0.1)		
TIMP4		1.71->0.0 (0.0)		1.19->5.28 (4.4)
TMEM154	1.43->0.17 (0.1)	1.24->3.77 (3.0)		

Table 3.3. Correlation between RNASeq and Whole mount *in situ* hybridization experiments. The results presented in this table are from genes validated using WISH by collaborates in the Karen Sears lab.
+ indicates correlation between RNASeq and WISH expression data
? indicates poor correlation between RNASeq and WISH expression data

		Mouse	Bat	Opossum	Pig
Fore limb bud	*Evx1*				
	a13	+			
	a11	?			
	Lnp	+			
	Evx2	+			
	d13	+			
	d12	+			
	d11	+			
	d10	+			
Fore limb paddle	*Evx1*	+	+		
	a13	+		+	+
	a11	+		+	+
	Lnp	+		+	
	Evx2	+		+	
	d13	+		+	
	d12	+		+	
	d11	+		?	
	d10	?			
Hind limb bud	*Evx1*	+		+	
	a13	+		+	
	a11	+		+	
	Lnp	+		+	
	Evx2	+		+	
	d13	+		+	
	d12	+		+	
	d11	+		+	
	d10	?		+	
Hind limb paddle	*Evx1*	?			
	a13	+			
	a11	+		+	
	Lnp	?			
	Evx2	+			
	d13	+		+	
	d12	+		+	
	d11	+		+	
	d10	+			

Table 3.4. Transcription factor-enhancer pairs with lineage specific changes near limb development genes.

Transcription Factor	Relative location of enhancer to Known Limb Development Genes	Enhancer Coordinates in Mouse Genome	Enhancer	Limb Development Gene(s)
MEF2D	Overlap with gene of interest	14:28523050-28523575	Chromatin Mark (H3k27ac) at Wanek Stage 3	wnt5a
NP_032283.3	64274 bp 5' (Upstream)	2:93746275-93749125	Chromatin Mark (H3k27ac) at Wanek Stage 3	ext2
NP_032283.3	64274 bp 5' (Upstream)	2:93746275-93749125	Chromatin Mark (H3k27ac) at Wanek Stage 3	ext2
FOSB	81611 bp 5' (Upstream)	11:96111550-96112750	Chromatin Mark (H3k27ac) at Wanek Stage 3	hoxb1-hoxb13
BARHL1	69499 bp 5' (Upstream)	2:93743175-93743900	Chromatin Mark (H3k27ac) at Wanek Stage 3	ext2
NR2C2	30021 bp 5' (Upstream)	2:74490016-74492038	Known Limb Enhancer: Global Control Region Conserved Sequence A	hoxd1-hoxd13-evx2-lnp
SIX2	Overlap with gene of interest	14:28518450-28519050	Chromatin Mark (H3k27ac) at Wanek Stage 6	wnt5a
MEF2D	Overlap with gene of interest	14:28523050-28523575	Chromatin Mark (H3k27ac) at Wanek Stage 3	wnt5a
KLF6	78323 bp 3' (Downstream)	15:103115175-103116575	Chromatin Mark (H3k27ac) at Wanek Stage 3	hoxc4-hoxc13
ZFP523	Overlap with gene of interest	6:52156000-52158100	Chromatin Mark (H3k27ac) at Wanek Stage 3	hoxa1-hoxa13-evx1
THRB	Overlap with gene of interest	14:28518450-28519050	Chromatin Mark (H3k27ac) at Wanek Stage 6	wnt5a
DLX6	Overlap with gene of interest	15:103007525-103010725	Chromatin Mark (H3k27ac) at Wanek Stage 3	hoxc4-hoxc13
LIN54	Overlap with gene of interest	14:28523050-28523575	Chromatin Mark (H3k27ac) at Wanek Stage 3	wnt5a
RFX5	Overlap with gene of interest	6:52317575-52318275	Chromatin Mark (H3k27ac) at Wanek Stage 3	hoxa1-hoxa13-evx1
CREM	Overlap with gene of interest	15:102949700-102960300	Chromatin Mark (H3k27ac) at Wanek Stage 3	hoxc4-hoxc13
BATF3	Overlap with gene of interest	11:96205925-96208000	Chromatin Mark (H3k27ac) at Wanek Stage 3	hoxb1-hoxb13

Table 3.4 (cont.) Transcription factor-enhancer pairs with lineage specific changes near limb development genes.

Transcription Factor	Relative location of enhancer to Known Limb Development Genes	Enhancer Coordinates in Mouse Genome	Enhancer	Limb Development Gene(s)
ZKSCAN5	69499 bp 5' (Upstream)	2:93743175-93743900	Chromatin Mark (H3k27ac) at Wanek Stage 3	ext2
SMARCC2	Overlap with gene of interest	6:52216750-52217000	Chromatin Mark (H3k27ac) at Wanek Stage 6	hoxa1-hoxa13-evx1
FOXA1	Overlap with gene of interest	6:52259650-52260275	Chromatin Mark (H3k27ac) at Wanek Stage 3	hoxa1-hoxa13-evx1
ENSMUSG00000079994	Overlap with gene of interest	14:28518450-28519050	Chromatin Mark (H3k27ac) at Wanek Stage 6	wnt5a
RUNX2	38406 bp 5' (Upstream)	15:102881000-102882725	Chromatin Mark (H3k27ac) at Wanek Stage 3	hoxc4-hoxc13
ZBTB12	40948 bp 5' (Upstream)	14:28464000-28464525	Chromatin Mark (H3k27ac) at Wanek Stage 6	wnt5a
FOXC1	Overlap with gene of interest	14:28518450-28519050	Chromatin Mark (H3k27ac) at Wanek Stage 6	wnt5a
ARID3C	71223 bp 3' (Downstream)	15:103108075-103109675	Chromatin Mark (H3k27ac) at Wanek Stage 3	hoxc4-hoxc13
MEF2D	Overlap with gene of interest	14:28523050-28523575	Chromatin Mark (H3k27ac) at Wanek Stage 3	wnt5a

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410.

Andrey G, Montavon T, Mascrez B, *et al.* (2013) A switch between topological domains underlies HoxD genes collinearity in mouse limbs. *Science* **340**, 1234167.

Avila-Arcos MC, Ho SYW, Ishida Y, *et al.* (2013) One Hundred Twenty Years of Koala Retrovirus Evolution Determined from Museum Skins. *Molecular Biology and Evolution* **30**, 299-304.

Ballandras-Colas A, Naraharisetty H, Li X, Serrao E, Engelman A (2013) Biochemical Characterization of Novel Retroviral Integrase Proteins. *Plos One* **8**, 9.

Blanchette M, Kent WJ, Riemer C, *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research* **14**, 708-715.

Bromham L (2002) The human zoo: endogenous retroviruses in the human genome. *Trends in Ecology & Evolution* **17**, 91-97.

Camacho C, Coulouris G, Avagyan V, *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421.

Chambers GK, MacAvoy ES (2000) Microsatellites: consensus and controversy. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* **126**, 455-476.

Cock PJA, Antao T, Chang JT, *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422-1423.

Coffin JM (2004) Evolution of retroviruses: Fossils in our DNA. *Proceedings of the American Philosophical Society* **148**, 264-280.

Cooper LN, Cretekos CJ, Sears KE (2012) The evolution and development of mammalian flight. *Wiley Interdisciplinary Reviews: Developmental Biology* **1**, 773-779.

Cooper LN, Sears KE (2013) Bat Evolution, Ecology, and Conservation. 3-20.

Cotney J, Leng J, Yin J, *et al.* (2013) The Evolution of Lineage-Specific Regulatory Activities in the Human Embryonic Limb. *Cell* **154**, 185-196.

Craigie R, Bushman FD (2012) HIV DNA Integration. *Cold Spring Harbor Perspectives in Medicine* **2**, 18.

Cretekos CJ, Wang Y, Green ED, *et al.* (2008) Regulatory divergence modifies limb length between mammals. *Genes and Development* **22**, 141-151.

Creyghton MP, Cheng AW, Welstead GG, *et al.* (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 21931-21936.

Cunningham F, Amode MR, Barrell D, *et al.* (2015) Ensembl 2015. *Nucleic Acids Res* **43**, D662-669.

Denner J (2007) Transspecies transmissions of retroviruses: New cases. *Virology* **369**, 229-233.

DeSalle R, Amato G (2004) The expansion of conservation genetics. *Nature Reviews Genetics* **5**, 702-712.

Driscoll CA, Menotti-Raymond M, Nelson G, Goldstein D, O'Brien SJ (2002) Genomic Microsatellites as evolutionary chronometers: A test in wild cats. *Genome Research* **12**, 414-423.

Duboule D (2007) The rise and fall of Hox gene clusters. *Development* **134**, 2549-2560.

Duque TSPC (2013) New insights into Cis - Regulatory Module Evolution using In - Silico Evolutionary.

Eggert LS, Eggert JA, Woodruff DS (2003) Estimating population sizes for elusive animals: the forest elephants of Kakum National Park, Ghana. *Molecular Ecology* **12**, 1389-1402.

Emslie R, Milliken T, Talukdar B (2013) African and Asian Rhinoceroses – Status, Conservation and Trade In: *A report from the IUCN Species Survival Commission (IUCN/SSC) African and Asian Rhino Specialist Groups and TRAFFIC to the CITES Secretariat pursuant to Resolution Conf. 9.14 (Rev. CoP15).* pp. CoP 16, doc 54.12.

Faircloth BC (2008) Msatcommander: detection of microsatellite repeat arrays and automated, locus-specific primer design. *Mol Ecol Resour* **8**, 92-94.

Ferrier DE, Holland PW (2001) Ancient origin of the Hox gene cluster. *Nature Reviews Genetics* **2**, 33-38.

Fiebig U, Hartmann MG, Bannert N, Kurth R, Denner J (2006) Transspecies transmission of the endogenous koala retrovirus. *Journal of Virology* **80**, 5651-5654.

Fujita PA, Rhead B, Zweig AS, *et al.* (2011) The UCSC Genome Browser database: update 2011. *Nucleic Acids Res* **39**, D876-882.

Giardine B, Riemer C, Hardison RC, *et al.* (2005) Galaxy: A platform for interactive large-scale genome analysis. *Genome Research* **15**, 1451-1455.

Grabherr MG, Haas BJ, Yassour M, *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**, 644-652.

Hanger JJ, Bromham LD, McKee JJ, O'Brien TM, Robinson WF (2000) The nucleotide sequence of koala (Phascolarctos cinereus) retrovirus: a novel type C endogenous virus related to gibbon ape leukemia virus. *Journal of Virology* **74**, 4264-4272.

Hayward JA, Tachedjian M, Cui J, *et al.* (2013) Identification of diverse full-length endogenous betaretroviruses in megabats and microbats. *Retrovirology* **10**, 19.

He X, Ling X, Sinha S (2009) Alignment and prediction of cis-regulatory modules based on a probabilistic model of evolution. *PLoS Computational Biology* **5**, e1000299.

Hedges S, Johnson A, Ahlering M, Tyson M, Eggert LS (2013) Accuracy, precision, and cost-effectiveness of conventional dung density and fecal DNA based survey methods to estimate Asian elephant (Elephas maximus) population size and structure. *Biological Conservation* **159**, 101-108.

Holland PW, Booth HA, Bruford EA (2007) Classification and nomenclature of all human homeobox genes. *BMC Biol* **5**, 47.

Houlden BA, Costello BH, Sharkey D, *et al.* (1999) Phylogeographic differentiation in the mitochondrial control region in the koala, Phascolarctos cinereus (Goldfuss 1817). *Molecular Ecology* **8**, 999-1011.

Ishida Y, Demeke Y, de Groot PJV, *et al.* (2011) Distinguishing Forest and Savanna African Elephants Using Short Nuclear DNA Sequences. *Journal of Heredity* **102**, 610-616.

Johnson WE, Coffin JM (1999) Constructing primate phylogenies front ancient retrovirus sequences. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 10254-10260.

Karolchik D, Barber GP, Casper J, *et al.* (2014) The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res* **42**, D764-D770.

Kellis M, Wold B, Snyder MP, *et al.* (2014) Defining functional DNA elements in the human genome. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 6131-6138.

Kim J, Cunningham R, James B, *et al.* (2010) Functional characterization of transcription factor motifs using cross-species comparison across large evolutionary distances. *PLoS Computational Biology* **6**, e1000652.

Kofler R, Schlotterer C, Lelley T (2007) SciRoKo: A new tool for whole genome microsatellite search and investigation. *Bioinformatics* **23**, 1683-1685.

Kohn MH, York EC, Kamradt Da, *et al.* (1999) Estimating population size by genotyping faeces. *Proceedings. Biological sciences / The Royal Society* **266**, 657-663.

Lamprecht B, Walter K, Kreher S, *et al.* (2010) Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. *Nature Medicine* **16**, 571-579.

Lander ES, Int Human Genome Sequencing C, Linton LM, *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357-U354.

Lehoczky Ja, Williams ME, Innis JW (2004) Conserved expression domains for genes upstream and within the HoxA and HoxD clusters suggests a long-range enhancer existed before cluster duplication. *Evolution & Development* **6**, 423-430.

Levinson G, Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular Biology and Evolution* **4**, 203-221.

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550.

Maas Sa, Fallon JF (2005) Single base pair change in the long-range Sonic hedgehog limb-specific enhancer is a genetic basis for preaxial polydactyly. *Developmental Dynamics* **232**, 345-348.

Meredith RW, Westerman M, Springer MS (2009) A phylogeny and timescale for the living genera of kangaroos and kin (Macropodiformes : Marsupialia) based on nuclear DNA sequences. *Australian Journal of Zoology* **56**, 395-410.

Mi S, Lee X, Li XP, *et al.* (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* **403**, 785-789.

Montavon T, Soshnikova N, Mascrez B, *et al.* (2011) A regulatory archipelago controls hox genes transcription in digits. *Cell* **147**, 1132-1145.

Okello JB, Wittemyer G, Rasmussen HB, *et al.* (2005) Noninvasive genotyping and Mendelian analysis of microsatellites in African savannah elephants. *Journal of Heredity* **96**, 679-687.

Portales-Casamar E, Thongjuea S, Kwon AT, *et al.* (2009) JASPAR 2010: The greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res* **38**, 105-110.

Reddy PS, Mahanty S, Kaul T, *et al.* (2008) A high-throughput genome-walking method and its use for cloning unknown flanking sequences. *Analytical Biochemistry* **381**, 248-253.

Renfree MB, Papenfuss AT, Deakin JE, *et al.* (2011) Genome sequence of an Australian kangaroo, Macropus eugenii, provides insight into the evolution of mammalian reproduction and development. *Genome Biology* **12**, 25.

Richardson MK, Gobes SM, van Leeuwen AC, *et al.* (2009) Heterochrony in limb evolution: developmental mechanisms and natural selection. *Journal of experimental zoology. Part B, Molecular and developmental evolution.* **312**, 639-664.

Roca AL, Peacon-Slattery J, O'Brien SJ (2004) Genomically intact endogenous feline leukemia viruses of recent origin. *Journal of Virology* **78**, 4370-4375.

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132**, 365-386.

Ruiz-Rodriguez CT, Ishida Y, Greenwood AD, Roca AL (2014) Development of 14 microsatellite markers in the Queensland koala (Phascolarctos cinereus adustus) using next generation sequencing technology. *Conservation Genetics Resources* **6**, 429-431.

Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* **32**, 91D-94.

Schneider I, Aneas I, Gehrke AR, *et al.* (2011) Appendage expression driven by the Hoxd Global Control Region is an ancient gnathostome feature. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 12782-12786.

Sears KE, Bormet AK, Rockwell A, *et al.* (2011) Developmental basis of mammalian digit reduction: a case study in pigs. *Evolution & Development* **13**, 533-541.

Selkoe KA, Toonen RJ (2006) Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters* **9**, 615-629.

Shimode S, Nakagawa S, Yoshikawa R, Shojima T, Miyazawa T (2014) Heterogeneity of koala retrovirus isolates. *Febs Letters* **588**, 41-46.

Shojima T, Hoshino S, Abe M, *et al.* (2013a) Construction and Characterization of an Infectious Molecular Clone of Koala Retrovirus. *Journal of Virology* **87**, 5081-5088.

Shojima T, Yoshikawa R, Hoshino S, *et al.* (2013b) Identification of a Novel Subgroup of Koala Retrovirus from Koalas in Japanese Zoos. *Journal of Virology* **87**, 9943-9948.

Sievers F, Wilm A, Dineen D, *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7**, 539.

Simmons GS, Young PR, Hanger JJ, *et al.* (2012) Prevalence of koala retrovirus in geographically diverse populations in Australia. *Australian Veterinary Journal* **90**, 404-409.

Sinha S (2007) PhyME: a software tool for finding motifs in sets of orthologous sequences. *Methods Mol Biol* **395**, 309-318.

Sinha S, van Nimwegen E, Siggia ED (2003) A probabilistic method to detect regulatory modules. *Bioinformatics* **19 Suppl 1**, i292-i301.

Smit A, Hubley R, P G (2013-2015) *RepeatMasker Open-4.0.* http://www.repeatmasker.org

Spitz F, Gonzalez F, Duboule D (2003) A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. *Cell* **113**, 405-417.

Stern DL (2000) Evolutionary developmental biology and the problem of variation. *Evolution* **54**, 1079-1091.

Stormo GD, Schneider TD, Gold L, Ehrenfeucht A (1982) Use of the 'perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res* **10**, 2997-3011.

Stoye JP (2006) Koala retrovirus: a genome invasion in real time. *Genome Biology* **7**, 3.

Stoye JP (2012) Studies of endogenous retroviruses reveal a continuing evolutionary saga. *Nature Reviews Microbiology* **10**, 395-406.

Taberlet P, Luikart G, Waits LP (1999) Noninvasive genetic sampling: Look before you leap. *Trends in Ecology & Evolution* **14**, 323-327.

Tarlinton R, Meers J, Young P (2008) Biology and evolution of the endogenous koala retrovirus. *Cellular and Molecular Life Sciences* **65**, 3413-3421.

Tarlinton RE, Meers J, Young PR (2006) Retroviral invasion of the koala genome. *Nature* **442**, 79-81.

Taylor AC, Graves JAM, Murray ND, Sherwin WB (1991) Conservation genetics of the koala (*Phascolarctos cinereus*). II. Limited variability in minisatellite DNA sequences. *Biochemical Genetics* **29**, 355-363.

Thitaram C, Thongtip N, Somgird C, *et al.* (2008) Evaluation and selection of microsatellite markers for an identification and parentage test of Asian elephants (Elephas maximus). *Conservation Genetics* **9**, 921-925.

Thurston MI, Field D (2005) *Msatfinder: detection and characterisation of microsatellites.* http://www.genomics.ceh.ac.uk/msatfinder/

Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111.

Untergasser A, Cutcutache I, Koressaar T, *et al.* (2012) Primer3--new *capabilities* and interfaces. *Nucleic Acids Res* **40**, e115.

Vokes Sa, Ji H, Wong WH, McMahon AP (2008) A genome-scale analysis of the cis-regulatory circuitry underlying sonic hedgehog-mediated patterning of the mammalian limb. *Genes and Development* **22**, 2651-2663.

Wanek N, Muneoka K, Holler-Dinsmore G, Burton R, Bryant SV (1989) A staging system for mouse limb development. *Journal of Experimental Zoology* **249**, 41-49.

Wasser SK, Brown L, Mailand C, *et al.* (2015) CONSERVATION. Genetic assignment of large seizures of elephant ivory reveals Africa's major poaching hotspots. *Science* **349**, 84-87.

Weirauch Matthew T, Yang A, Albu M, *et al.* (2014) Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell* **158**, 1431-1443.

Xu WQ, Stadler CK, Gorman K, *et al.* (2013) An exogenous retrovirus isolated from koalas with malignant neoplasias in a US zoo. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 11547-11552.