Development of Bioinformatic Tools for Retroviral Analysis from High Throughput Sequence Data

Dissertation

zur Erlangung des Grades eines Doktors der Naturwissenschaften (Dr. rer. nat.)

am Fachbereich Mathematik und Informatik der
Freien Universität Berlin vorgelegt von

Ulrike Löber

Berlin 2019

Datum der Disputation: 18.06.2019
Erstgutachter:
Prof. Dr. Knut Reinert, Freie Universität Berlin, Deutschland
Zweitgutachter:
Prof. Dr. Alex Greenwood, Freie Universität Berlin, Deutschland

“Nothing in Biology Makes Sense Except in the Light of Evolution”
Theodosius Dobzhansky (March 1973)

Preface

Comprehensive profiling of retroviral integration sites using target enrichment methods from historical koala samples without an assembled reference genome The work presented in Chapter 2 was published in the journal PeerJ under peer review. As part of this project, I developed the bioinformatics pipeline to detect viral insertion sites from highly degraded DNA. Furthermore, I helped conceive and design the experiments, analyzed the data, co-wrote the paper, prepared figures and tables and reviewed drafts of the paper.

> Cui, P., **Löber, U.**, et al. (2016). Comprehensive profiling of retroviral integration sites using target enrichment methods from historical koala samples without an assembled reference genome. PeerJ 4, e1847 [1].

Degradation and remobilization of endogenous retroviruses by recombination during the earliest stages of a germline invasion The work presented in Chapter 4 was published in the journal Proceedings of the National Academy of Sciences (PNAS) under peer review. In the context of this project, I helped conceive and design the experiments, analyzed the data, co-wrote the paper, prepared figures and tables and reviewed drafts of the paper. `https://doi.org/10.1073/pnas.1807598115`

> **Löber, U.**, et al. (2018). Degradation and remobilization of endogenous retroviruses by recombination during the earliest stages of a germ-line invasion. PNAS 201807598. [2]

SIP: A Sonication Based Inverse PCR Methodology for the Genome-Scale Analysis of Retroviral Integration Sites and Provirus Characterization The work presented in Chapter 3 is in preparation for submission. As part of this project, I developed the bioinformatics pipeline to detect viral insertion sites from PacBio reads. Furthermore, I analyzed the data, prepared figures and tables, and wrote and reviewed drafts of the paper.

> Alquezar-Planas, D.E., **Löber, U.**, Cui, P., Quedenau, C., Chen, W. and Greenwood, A.D. (submitted). SIP: DNA Sonication Inverse PCR for Genome-Scale Analysis Integration Sites. [3]

Abstract

For hundreds of millions of years, retroviruses have been integrating into genomes of vertebrates [4]. This thesis contributes to the development of new methods for retrieval, characterization and the comparison of viruses that have integrated into the genome (endogenous retroviruses, or ERVs) and their integration sites in host genomes. The koala retrovirus is an outstanding study subject since it is currently in the transition from an exogenous to an endogenous retrovirus.

In the past decades, high-throughput sequencing (HTS) has allowed scientists to investigate genomic data at high coverage and low costs. However, the development of new sequencing technologies facilitated the production of vast amounts of data. The analysis bottleneck has shifted from data production to the analysis of so-called "big data". In consequence, new algorithms and pipelines need to be established to process biological data. Solutions for automated handling of short-read HTS data exist for many problems and can be improved and extended. Recent improvements in HTS resulting in longer sequence fragments have helped solve problems connected to short-read sequencing but produced new challenges for genomics data processing.

In this thesis, I present pipelines to comprehensively profile endogenous retroviruses from short-read HTS data for museum koala samples (ancient DNA) and describe a new method to amplify retroviral integration sites facilitating long-read HTS. The thesis is divided into five sections. In the first part, I describe the biological problem, the evolution of sequencing technologies, resulting in information technology problems and proposed solutions (chapter 1). In the second chapter, I present a comparison of three different target enrichment techniques to retrieve retroviral integration sites from museum koala samples. The computational pipeline I developed for this purpose is presented. In chapter 3 I describe a method (sonication inverse polymerase chain reaction) for target enrichment of long sequence fragments to exploit the capacities of third-generation sequencing technologies. An analysis pipeline for the processing of sonication inverse PCR products was established. Moreover, remaining problems resulting from artificial read structures are discussed. In chapter 4 the method described in chapter 3 was used to profile koala retrovirus integrations. The striking discovery of a new retroviral recombinant in koalas is reported. Finally, I discuss our findings and compare short- and long-read HTS technologies. An outlook for further applications and remaining computational problems is outlined.

Overall, this thesis contributes to the automated computational processing of HTS data from target enrichment techniques to profile endogenous retroviruses in host genomes.

Zusammenfassung

Seit Millionen von Jahren nisten sich Retroviren in den Genomen von Wirbeltieren ein [4]. Die hier vorliegende Dissertation zeigt neue Methoden zur Gewinnung und Analyse genomischer Sequenzen retroviralen Ursprungs und zum Vergleich retroviraler Integrationsstellen in Wirtsgenomen auf. Die Analysen werden am Beispiel des Koalaretroviruses durchgeführt, da sich dieser im Übergang vom exogenen zum endogenen Retrovirus befindet. Die Erforschung des Endogenisierungsprozesses von Retroviren kann so untersucht werden.

Die in den letzten Jahrzehnten erzielten Fortschritte in der Hochdurchsatzsequenzierung machen die Erzeugung von Sequenzierdaten mit relativ geringem Kostenaufwand möglich. Die Entwicklung in der Sequenzierungstechnologie stellt durch die Produktion beträchtlicher Datenmengen und der damit verbundenen notwendigen Verarbeitung eine große Herausforderung dar. Während zuvor labortechnische Aspekte die Kapazitäten von aufwendigen Sequenzanalysen limitierten, sind heute die Datenspeicherung und -verarbeitung die begrenzenden Faktoren. In Folge dessen gewinnt die Entwicklung neuer Algorithmen und Pipelines zur Verarbeitung großer Mengen von biologischen Sequenzdaten an Bedeutung. Für Hochdurchsatzsequenzierungsdaten mit kurzen Fragmentlängen sind bereits viele informationstechnologische Problemstellungen bearbeitet und gelöst worden. Es besteht der Bedarf an der Lösung verbleibender Probleme und der Optimierung bereits entwickelter Algorithmen und Software. Durch die Evolution der Hochdurchsatzsequenzierung entstehen Rohdaten aus deutlich längeren Segmenten, welche häufig nicht oder nicht zufriedenstellend durch existierende Software verarbeitet werden können.

In der vorliegenden Arbeit stelle ich verschiedene Pipelines zur Verarbeitung unterschiedlicher Hochdurchsatzsequenzierungsdaten vor. Im ersten Kapitel gebe ich eine Einführung in die biologische Fragestellung, beschreibe verschiedene Sequenzierungsmethoden und deren Entwicklung, sowie existierende informationstechnologische Lösungen. Anschließend stelle ich in Kapitel 2 eine Arbeit vor, bei welcher ich eine Pipeline entwickelt habe, um endogene Retroviren aus Koalaexponaten mit Hilfe von kurzen Sequenzfragmenten zu vergleichen. Kapitel 3 behandelt eine neue Methode zur zielgerichteten Amplifizierung längerer Sequenzfragmente, um die Vorteile neuerer Hochdurchsatzsequenzierungstechniken auszunutzen. Auf der Grundlage dieser Methode habe ich eine Pipeline entwickelt, um endogene Koalaretroviren in einem 2014 verstorbenen Zookoala zu klassifizieren, wobei ein neuer rekombinanter Virus entdeckt und charakterisiert werden konnte. Die Vor- und Nachteile verschiedener Hochdurchsatzsequenzierungstechnologien, sowie ungelöste Probleme und ein Ausblick beinhaltet die Diskussion in Kapitel 5.

Zusammengefasst konzentriert sich die vorliegende Arbeit auf verschiedene Methoden zur automatisierten informationstechnologischen Verarbeitung von Hochdurchsatzdaten, zum Vergleich endogener Retroviren in Wirtsgenomen unter der Nutzung verschiedener Anreicherungs- und Sequenziertechniken.

Acknowledgment

This thesis is based on research conducted at the Leibniz Institute for Zoo and Wildlife Research. I would like to show my gratitude to Alex Greenwood for giving me the opportunity to work on these exciting projects, his advice, his patience, and his contagious enthusiasm. The fruitful discussions with you were enlightening, fascinating and open minded. The advice given by Knut Reinert has been a great help in conducting this thesis and evaluating technical approaches.

I would like to pay my regards to Anisha Dayaram and David Alquezar for all the experiments they have performed and for discussions. Sanatana, Daniela, John, Hanna, Sonia, Renata, Marcella, Anisha, David, and Niccolo, you were great company during the last years, it would not have been the same without you; thanks for the laughter, the ice-skating action, the moments we shared, the tears and everything which made the last years unforgettable. Thanks, are also due to Pin Cui, for performing experiments. I would like to express my gratitude to the people at BeGenDiv for hosting me once a week. I would like to thank my fellow doctoral student Peter Seeber for his feedback and friendship.

Special thanks to my high school teacher Ms. Hermann who awoke my interest in biology. I extend my gratitude to Stefanie Hartmann: you are the reason why I started studying bioinformatics. Your guidance is unrivaled, thank you for leading me the right way. I appreciate the feedback offered by Sofia Forslund and want to thank you for your patience.

Last but not least; Dear Mom and Dad, there are no words describing how thankful I am that you can experience this special moment with me. I know that this is a gift and instead of listing everything you mean to me or things I am grateful for, I just want to state that I love you and hope I can share some more time with you. Thanks to Sophie for your patience and backing during the last years, you encouraged me whenever I was close to surrender.

The financial support of the Forschungsverbund Berlin is also gratefully acknowledged.

Contents

1 Introduction **1**
1.1 Retroviridae . . . 1
1.1.1 Endogenous Retroviruses . . . 2
1.1.2 Koala Retrovirus . . . 3
1.1.3 Hypothesis: Viral Insertion and Defense Mechanisms . . . 3
1.2 Sequencing Technologies . . . 4
1.2.1 Next/Second-Generation Sequencing . . . 4
1.2.2 Third-Generation Sequencing . . . 5
1.3 Bioinformatics Analysis of Sequencing Data . . . 7
1.3.1 Pairwise Sequence Alignment . . . 8
1.3.2 Multiple Sequence Alignment . . . 10
1.3.3 Sequence Mapping . . . 11
1.3.4 Sequence Clustering . . . 12
1.3.5 Tools and Pipelines for Retrieval of Retroviral Integration Sites . 13
1.4 Objective . . . 14
1.4.1 Retroviral Integration Sites . . . 14

2 Investigation of Koala Retrovirus in Museum Koala Samples **16**
2.1 Characterizing Viral Integration Sites from Ancient DNA . . . 16
2.1.1 Comparison of Three Targeted Sequencing Methods . . . 17
2.1.2 Detection of the Koala Retrovirus from Illumina Short-Reads . . . 19
2.2 Results . . . 22
2.2.1 Target Enrichment Techniques for Integration Site Retrieval from Ancient DNA . . . 22
2.2.2 Comprehensive Profiling of Koala Retrovirus (KoRV) Integration Sites . . . 26
2.3 Discussion . . . 28

3 Sonication Inverse Polymerase Chain Reaction (SiP) **30**
3.0.1 Laboratory Pipeline . . . 31
3.1 Methods . . . 32
3.1.1 Bioinformatic Analysis . . . 37
3.2 Results . . . 37
3.2.1 Development and Testing of SiP . . . 37
3.2.2 Evaluating SiP's Library Length Distribution, KoRV Sequence Enrichment and Off Target Enrichment . . . 39
3.2.3 SiP Blunt-End Adapter Ligation Efficiency and Blunt-End 454 DNA Library Adapter Ligation Experiment . . . 45

3.2.4 SiP Structure Variations . . . 48
3.3 Discussion . . . 48
3.3.1 Transposition of Inverse Polymerase Chain Reaction Products . . 51
3.3.2 Assembling Complete Proviruses at Genomic Loci . . . 52

4 Investigation of Koala Retrovirus in Modern Samples 54
4.1 Methods . . . 55
4.1.1 Koala Samples, PCR, and Sequencing . . . 56
4.1.2 Inverse PCR and PacBio Sequencing of "Bilyarra" to Determine KoRV and recKoRV Sequences . . . 57
4.1.3 Bioinformatics Analysis . . . 59
4.2 Results . . . 62
4.2.1 The Advantages of Long Read Sequence Technology for Retroviral Analysis . . . 62
4.2.2 Discovery of recKoRV1 [2] . . . 63
4.2.3 Other Recombinants Between KoRV and PhER . . . 64
4.2.4 Comparison of LTRs and Integration Sites Among Koala Genomes 70
4.2.5 Recombination Breakpoint Distributions Among Koala Populations 72
4.2.6 An Extended Analysis of the Geographic Distribution of recKoRV1 73
4.3 Discussion . . . 73
4.3.1 recKoRV Classification . . . 76

5 Conclusions 78
5.1 Resume . . . 78
5.2 Perspectives/Outlook . . . 79

A Appendix 81
A.1 Supplementary Figures and Tables . . . 101

List of Figures

1.1 Structure of an Endogenous Retrovirus 3
1.2 PacBio Single Molecule Real Time Sequencing 7
1.3 Schematic Hidden Markov Model 13

2.1 Experimental Workflow for Three Target Enrichment Techniques [1] . . . 18
2.2 Primer Positions and LTR Structure for the Identification of KoRV Integration Sites [1] 20
2.3 Bioinformatics Pipeline for Targeted ERVs from Museum Samples 21
2.4 Sequence Length Distribution of Three Target Enrichment Techniques [1] 23
2.5 Expected Product Distribution of Target Enrichment Techniques 24
2.6 Overview of Integration Sites from Ten Individuals Across Target Enrichment Techniques 25
2.7 Number of Individuals from this Study per Shared Integration Site . . . 27

3.1 Sonication Inverse PCR Scheme [3] 33
3.2 Inverse PCR Primer 34
3.3 Experimental Workflow of Sonication Inverse PCR [3] 38
3.4 Reads of Insert (ROI) Sequence Length Distribution [3] 40
3.5 Adapter Long Off-Targets 42
3.6 Adapter Short Off Targets 43
3.7 Non-Adapter Long Off Targets 44
3.8 Non-Adapter Short Off Targets 45
3.9 Blunt-End 454 DNA Library Adapter Ligation Experimental Workflow [3] 47
3.10 Identified Structural Variants of SiP Sequences [3] 49
3.11 Expected Structure of Inverse PCR Reads 52
3.12 KoRV Integration Site Shared with Koala Reference Genome (KoRV 35) 53

4.1 Example of Read Mapping from Captured Forward KoRV Integration Site 63
4.2 Recombination Breakpoints of KoRV and PhER [2] 64
4.3 Mapping KoRV Positive Reads to Breakpoint “Birke” bp12 66
4.4 Mapping KoRV Positive Reads to Breakpoint “Birke” bp13 66
4.5 Mapping KoRV Positive Reads to Breakpoint “Birke” bp14 67
4.6 Mapping KoRV Positive Reads to Breakpoint “Birke” bp16 67
4.7 Mapping KoRV Positive Reads to Breakpoint “Birke” bp17 67
4.8 Recombination Breakpoints and Integration Sites Across Koalas [2] . . . 68
4.9 Relationships Among KoRV LTR Sequences [2] 71
4.10 Prevalence of recKoRV1 in KoRV-positive and KoRV-negative Koalas Across Australia 74
4.11 Mapping Result for Scaffold 3 recKoRV Reverse 77

A.1 Scaf14 PCR Products Sanger Sequencing 107
A.2 Scaf21 PCR Products Sanger Sequencing 108
A.3 Scaf24 PCR Products Sanger Sequencing 109
A.4 Scaf37 16563123 PCR Products Sanger Sequencing 110
A.5 Scaf37 17582840 PCR Products Sanger Sequencing 111
A.6 Scaf69 PCR Products Sanger Sequencing 112
A.7 Scaf79 PCR Products Sanger Sequencing 113
A.8 Scaf83 PCR Products Sanger Sequencing 114
A.9 Scaf96 PCR Products Sanger Sequencing 115
A.10 Scaf173 PCR Products Sanger Sequencing 116
A.11 Scaf275 PCR Products Sanger Sequencing 117

List of Tables

2.1 Museum Koala Sample Information [1] 17
2.2 Sequencing Statistics of Target Enrichment Techniques[1] 19
2.3 megaBLAST and TRIBE-MCL Parameters for Sequence Clustering [1] . 21
2.4 Sequence Statistics for Integration Site Detection from aDNA [1] 25
2.5 Shared Integration Sites Across Studies 28

3.1 List of Primers and Oligonucleotides [3] 35
3.2 Sequencing Statistics SiP PacBio 36
3.3 KoRV Sequence Enrichment Using SiP and PacBio RSII Sequencing . . . 41
3.4 Adapter Counts Across Four SiP Datasets and Blunt-End 454 DNA Library (with Illumina Adapters) Dataset 46

4.1 Koala Samples and Datasets Utilized [2] 58
4.2 recKoRV1 Classification Summary 60
4.3 Reads Spanning Breakpoints bp12, bp13, bp16, bp17 65
4.4 recKoRVs Identified in the Current Study and their Distribution in Modern and Historical Koala Genomes [2] 69
4.5 Shared Target Site Duplications “Bilyarra” 72

A.1 KoRV Forward Integration Sites Identified in Bilyarra and Their Genomic Locations Relative to the Koala Reference Genome 101
A.2 KoRV Reverse Integration Sites and Integration Sites of Unknown Orientation Identified in Bilyarra and Their Genomic Locations Relative to the Koala Reference Genome 102
A.3 recKoRV Integration Sites Identified in Bilyarra and Their Genomic Locations Relative to the Koala Reference Genome 103
A.4 Alignment of PCR Sanger Products to Viral Domains, Breakpoints, and Insertion Sites 104
A.5 Length of Sanger Sequenced PCR Products 105

Acronyms

aDNA ancient DNA

ENV envelope

env envelope gene

ERV endogenous retrovirus

GAG group-specific antigen

gag group-specific antigen gene

gDNA genomic DNA

HC Hybridization Capture

HTS high-throughput sequencing

iPCR inverse polymerase chain reaction

KoRV koala retrovirus

LTR long terminal repeat

MSA multiple sequence alignment

ORF open reading frame

PCR polymerase chain reaction

PEC Primer Extension Capture

PhER Phascolarctos endogenous retroelement

POL polymerase

pol polymerase gene

recKoRV recombinant koala retrovirus

ROI read of insert

SPEX Single Primer Extension

Chapter 1

Introduction

For hundreds of millions of years, retroviruses have been integrating into vertebrate genomes [4]. This thesis contributes to the development of new methods for retrieval, characterization, and the comparison of endogenous retroviruses (ERVs) and their integration sites in host genomes. The koala retrovirus (KoRV) is an outstanding study subject, since it is currently in the transition from an exogenous (infectious) to an endogenous (genomic trait) retrovirus and is one of the only mammalian retroviruses that is in such an early stage of genomic invasion. Many questions about retroviral endogenization are still open. Investigating an endogenous retrovirus in the earliest stages of genome invasion may provide insights into the underlying mechanisms of endogenization, including the interdependencies of host and virus.

In 1983, the human immunodeficiency virus (HIV), causing the acquired immunodeficiency syndrome, was described. HIV causes nausea, vomiting, persistent diarrhea, chronic fatigue, rapid weight loss, cough and shortness of breath, recurring fever, chills and night sweats, lesions in the mouth or nose, on the genitals or under the skin. Theodor Bestor proposed that HIV might endogenize into the human genome within a lifetime [5].

1.1 Retroviridae

Retroviruses are RNA viruses with a DNA intermediate that integrates into the genome of host cells. The host cell subsequently acts as a reservoir for new virus particles. Retrovirus stands for “Reverse Transcriptase Oncovirus”. Thus, these viruses can reverse-transcribe their RNA to DNA. Using an integrase gene, the virus to insert its genetic information into the host’s genome. Non-infectious particles with the same integration mechanism are called retrotransposons. Retroviruses and retrotransposons form the group of retroelements.

Retroviruses are classified in eleven different genera: Alpharetroviruses, Betaretroviruses, Gammaretroviruses, Deltaretroviruses, Epsilonretroviruses, Lentiviruses, Bovispumaviruses, Equispumaviruses, Felispumaviruses, Prosimiispumaviruses, and Simiispumaviruses [6]. Retroviruses infect different cell types, such as lymphocytes, T-cells or germ cells. Usually, the described mechanism leads to virus proliferation; new viral particles may infect other individuals of the same host species or hosts. Such horizontally transmitted viruses are exogenous retroviruses. Retroviruses can cause a variety of diseases like anemia, arthritis, cancer, mastitis, osteopetrosis, pneumonia, modest growth, immunosuppression resulting from atrophy of the bursa and thymus [7]. If a retrovirus infects a germline cell, the virus could be transmitted vertically, thus parents will pass the viral genetic information on to their offspring, which constitutes an ERV founder event.

Integrants are referred to as proviruses. For the sake of conciseness, proviruses flanking host genomic sequences are referred to as integration sites hereafter. During integration, single-strand gaps of the host-cell are repaired and result into a 4-10 bp duplication flanking the provirus, referred to as target site duplication. Retroviruses do not integrate randomly into host genomes [8]. It is statistically significant that integrations into transcription units, +/- 2 kbp from the transcription start sites, and +/- 2 kbp from CpG islands are favored over random insertions. Different retroviral genera show different insertion preferences [9].

The retroviral genome contains three protein-coding domains:

1. group-specific antigen (GAG); cleavage products are the major structural proteins of the virus core
2. polymerase (POL); cleavage products always include reverse transcriptase and integrase
3. envelope (ENV); cleavage products surface and transmembrane are the structural proteins of the viral envelope [10].

Two long terminal repeats (LTRs) flank the viral protein-coding sequences. Different processes like reinfection, further germline retrotransposition, negative selection, and genetic drift determine the abundance of retroelements in a gene pool.

1.1.1 Endogenous Retroviruses

Even though fundamental insights into retroviral integration were obtained within the last decade, the overall mechanism of retroviral integration remains unclear [11]. The following research sheds light on several processes for investigating ERVs and on recombination processes during the early evolution of ERVs. Approximately 8% of the human genome is of retroviral origin. Retroelements may either be advantageous for the host, lower its fitness, or have no effects on the host (neutral). A host infected by a retrovirus may be protected against infections by similar retroviruses. This phenomenon is called superinfection resistance [12]. A similar mechanism has been reported for ERVs. Transposition of defective endogenous retroelements to recently integrated viruses could harm viral proliferation and thus be advantageous for the host [13].

It has been shown, that high levels of reverse transcriptase derived from ERVs might play a positive role in the host defense mechanism against infections of non-retroviral RNA viruses [14]. Retroelements play an essential role in transcriptomic profiles [15], genetic variability, epigenetic gene regulation [5], embryogenesis [16] and during the development of the placenta in mammals [17, 18]. The ambiguous effects of retroviral invasion into host genomes, like malignancy and the capacity to produce infectious viruses on the one hand, positive and/or negative immune modulations for the host, increase of genetic variability on different levels, on the other hand, are still under investigation [19, 20, 21].

Once integrated in a host genome, ERVs are inherited as a Mendelian trait (structure shown in figure 1.1). In consequence, recombination, degradation, mutations and genetic drift affect the proliferation and the impact of ERVs in the host genome.

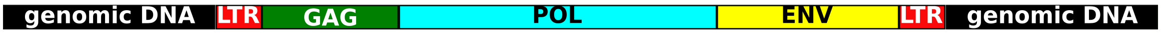


Figure 1.1: Structure of an Endogenous Retrovirus

Genomic DNA (black) is flanking the integrated endogenous retrovirus. Adjacent to genomic DNA are the duplicated long terminal repeats (LTRs) in red. Three protein-coding domains are present: the group-specific antigen (GAG) in green, the polymerase (POL) in blue and the envelope (ENV) domain in yellow.

1.1.2 Koala Retrovirus

The koala (*Phascolarctos cinereus* (Goldfuss, 1817)) is a solitary living species, ranging from New South Wales to South Australia. The International Union for Conservation of Nature Red List of threatened species categorizes the koala as vulnerable, while the population size is still decreasing [22].

An adult koala weighs between 4 and 15 kg, on average, whereas body size typically ranges from 60 to 85 cm. The diet of koalas consists almost exclusively of *Eucalyptus* leaves; therefore, koalas are confined to *Eucalyptus* forests. An average habitat size of 1.7 ha has been reported, and individuals rarely interact except in the breeding season when home ranges of males and females may overlap. Female koalas start breeding after the fourth year with a usual litter size of one, while the dominant males reproduce after the fifth year. Captive koalas can live up to 20 years [23]. The closest related extant marsupial species are wombats (*Vombatidae*) and kangaroos (*Macrops spec.*) [24]. My primary study subject, the male koala “Bilyarra” (Pci-SN241) from Tiergarten Schönbrunn in Vienna (Austria), was euthanized in July 2014, when he was 16 years old.

In 1988, KoRV was described for the first time to be associated with leukemia in koalas. The first case of leukemia in koalas was reported in the 1960s [25, 26]. Today it is assumed that KoRV infection results in neoplasia, causing lymphoma and leukemia, increases the prevalence of chlamydia infections and leads to immunomodulation [27, 28, 29]. KoRV is a Gammaretrovirus. Viruses most closely related to KoRV are the gibbon ape leukemia virus, the feline leukemia virus, and the porcine endogenous retrovirus [30, 31]. Three major subgroups of KoRV have been described, KoRV-A, KoRV-B/-J and the paraphyletic group KoRV-C/-D/-E/-F/-G/-H/-I [32]. According to the current state of science, KoRV-A is transmitted horizontally and vertically, whereas other subtypes like KoRV-B are exogenous and are only transmitted horizontally [32, 30, 33].

A gradient of infections with KoRV-A in koalas in eastern Australia was observed, with a prevalence of 100% in the North to 14.8% infected individuals on Kangaroo Island [34].

1.1.3 Hypothesis: Viral Insertion and Defense Mechanisms

Different defense mechanisms against the integration of ERVs are known, although incorporation of ERVs does not only have adverse effects for the host. Phenomena described as retroviral super-infection resistance, suggest that retroelements might act to block receptors against infections of related exogenous retroviruses [35, 12].

In my thesis, I hypothesize that recombination with other retroviral elements might play a role as a molecular defense mechanism against invading retroviruses that attempt to invade the germline. Furthermore, recombination, based on microhomologies with similar ERVs, impedes viral proliferation by introducing frameshifts in open reading frames (ORFs) or deletions. To test this hypothesis, I developed different methods, to examine ERV integrations in museum koala samples (chapter 2) and modern koala (chapter 3). A comparison and description of KoRV and recombinant koala retrovirus (recKoRV) integration sites in two modern koalas is outlined in chapter 4.

1.2 Sequencing Technologies

Oswald T. Avery, Colin M. MacLeod, and Maclyn McCarty found that genes and chromosomes consist of DNA, in 1944 [36]. The first complete nucleotide sequence was resolved in 1965 by Rober W. Holley and colleagues [37]. The team was able to determine 77 bp out of 80 bp long nucleotide sequence of alanine RNA, by purification of tRNAs from yeast, isolation of alanine RNA, different steps of digestion and ion-exchange chromatography. The experiments took more than five years. In 1977, methods to rapidly and accurately sequence DNA were developed. Sequencing was performed based on selective incorporation of radioactively or fluorescently labeled chain-terminating dideoxynucleotides during in vitro DNA replication, followed by size separation of fragments using gel electrophoresis [38, 39]. More than 20 years after James Watson and Francis Crick described the double helix structure of DNA, Sanger sequencing was the second technology which enabled scientists to investigate DNA sequences. Sanger sequencing required less handling of toxic chemicals and radioisotopes than the Maxam and Gilbert method, such that Sanger sequencing became the method of choice. First generation sequencing methods produce reads of up to 1 kbp length.

Since the late 2000's, different companies have developed various methods for high-throughput sequencing (HTS) at low costs, and though the produced reads were initially considerably shorter than Sanger sequencing reads, these next-generation techniques enabled scientists to perform massive parallel sequencing to decode complete genomes. Second-generation sequencing led to the so-called "genomics revolution". The third generation of sequencing techniques arose around 2010. Nanopore and Single Molecule Real Time Sequencing (SMRT) increased portability of sequencers, speed, read length and even empowered researchers to investigate epigenetis by detection of methylation. Nevertheless, until today second-generation sequencing such as Illumina sequencing persist due to the higher costs of long-molecule sequencing [40]. At the beginning of November 2018, Illumina, which is a leading company for second-generation sequencing, bought Pacific Biosciences, which was the leading third-generation sequencing company, for US$ 1.2B. It remains to be seen how this takeover will affect sequencing costs, technological improvements and sequencing evolution.

1.2.1 Next/Second-Generation Sequencing

There are different methods of next-generation sequencing, all of which have in common that they produce short sequence fragments at higher sequencing depth and lower costs than Sanger sequencing.

MALDI was the first system based on sequencing by mass spectrometry, developed in the late 80s. This method was able to detect methylation patterns and to do haplotype analysis [41]. Sequencing by hybridization was first described in 1998 [42]. Commercial systems are available from Affymetrix, NABsys and Complete Genomics Inc.. Currently, sequencing by hybridization and sequencing by mass spectrometry has been mostly replaced by other sequencing methods. Sequencing by ligation was introduced in 2005, when Shendure and colleagues described a method using emulsion PCR combined with paramagnetic beads, featuring high signal density, geometric uniformity, and robust feature separation [43]. SOLiD (Sequencing by Oligonucleotide Ligation and Detection) is a modern platform that produces shorter reads at less sequencing depth than the aforementioned pyrosequencing technologies but is cheaper.

Simultaneously, 454, later acquired by Roche, released a GS20 sequencing machine in 2005. The GS20 is based on sequencing by synthesis (bead amplification). In 2006 Ju and colleagues reported a method for sequencing by synthesis using a solid surface (DNA Chip) approach and a four-color fluorescent scanner to detect light emission of fluorescently tagged nucleotides in a polymerase-based synthesis [44]. Sequencing by synthesis with bridge amplification, was technically developed further by the company Solexa, later acquired by Illumina, and is currently the market leader for second-generation sequencing. While the cost of sequencing the human genome amounted to approximately US$ 1M in 2007, Illumina reduced the costs to US$ 4000 in 2011, and the costs further dropped below US$ 1000 per genome in 2014.

All these improvements made it possible to sequence genes at high depth, quantify rare transcripts, provide information about alternative splicing and determine single nucleotide polymorphisms. This progress paved the way for comprehensive evolutionary studies and various other applications [45]. All next-generation sequencing platforms have their strength and limitations. While Roche (454 GS Junior) sequencers produce the longest reads and create the most contiguous assemblies, Life Technologies (Ion Torrent PGM) sequencers have the smallest error rates and Illumina (MiSeq) machines produce the highest throughput [46]. Illumina, Roche, Life Technologies and others have developed different sequencers filling niches in molecular biology. All these sequencers have assets and drawbacks, creating a broad market of commercially available sequencing technologies.

While Sanger Sequencing was used to investigate the first human genome, next-generation sequencing has a brought range of applications at low costs.

Without HTS projects like the investigation of honeybee disappearance by gut metagenomics analysis or sequencing of the first Neanderthal genome could not have been completed [47]. Illumina is the leader in the next-generation sequencing industry. In chapter 2 I used an Illumina MiSeq to investigate the distribution of KoRV from museal samples and compare target enrichment techniques for viral insertion sites utilizing second-generation sequencers.

1.2.2 Third-Generation Sequencing

Third-generation sequencing was the next breakthrough in sequencing technology. In 2003 Ido Braslavsky, Benedict Hebert, Emil Kartalov, and Stephen R. Quake first described a method to obtain sequence information from single DNA molecules [48]. Therefore, PCR is not needed before sequencing, which shortens DNA preparation time for sequencing, and overcomes amplification and dephasing biases introduced by PCR. There are three long-fragment HTS platforms: nanopore-sequencing, advanced microscopy techniques direct imaging of DNA molecules, and single-molecule real-time sequencing (SMRT) by synthesis [49]. The signal is captured in real time, which means that the signal, regardless whether it is fluorescent (Pacific Biosciences) or electric current (Nanopore), is monitored during the enzymatic reaction of adding nucleotides in the complementary strand [50]. One of the most significant benefits is substantially increased read length; however, the error rate is generally higher compared to short-read platforms. These technological improvements have ushered in a new era of molecular biology and genetics. For the analysis of ERVs described in chapter 3 and chapter 4, I employed PacBios' RSII platform to inspect long stretches of host genomic DNA flanking viral integration sites. In brief, SMRT sequencing (PacBio) is based on the following principles:

1. ligate hairpin adapter on both ends of target double-stranded DNA → single-stranded circular DNA
2. load product on a chip termed SMRT cell
3. single unit called zero-mode waveguide, with immobilized single polymerase
4. polymerase binds to either of the hairpin adapters (replication start)
5. fluorescently labeled nucleotides emit light pulse when incorporated
6. light impulses are tracked as a movie (0.5-4 h)
7. single-stranded circular DNA can be sequenced multiples times (passes) to create circular consensus sequences (CCS) [referred to as reads of insert (ROIs)] with higher accuracy [51]

The principles of PacBio SMRT sequencing are shown in figure 1.2.

The maximum read length of the PacBio RSII is above 20 kbp, which enables scientists to overcome typical problems of next-generation sequencing, such as assembling low complexity or repetitive regions.

The technological evolution has yielded new problems of sequence processing, highlighted in section 1.3. Currently, bioinformatic processing of sequencing data often remains to be a bottleneck of genetic analysis, since hard disc capacities roughly double every year, the costs of sending a bit over optical networks halves every nine months, whereas next-generation sequencing capacities have doubled in less than every six months since 2004 [52].

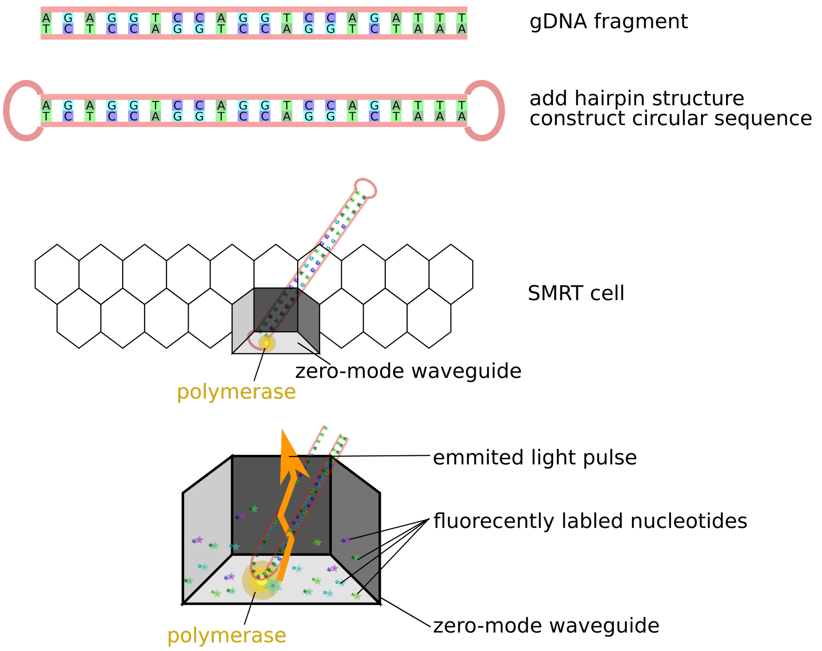

A G A G G T C C A G G T C C A G A T T T
T C T C C A G G T C C A G G T C T A A A
gDNA fragment
add hairpin structure
construct circular sequence
SMRT cell
zero-mode waveguide
polymerase
emmited light pulse
fluorecently labled nucleotides
zero-mode waveguide
polymerase

Figure 1.2: PacBio Single Molecule Real Time Sequencing

Hairpin adapters are ligated to fragmented DNA to produce circular sequences. Polymerase enzymes are attached to a sequencing matrix. One polymerase is immobilized in every single unit. All units together form a SMRT-Cell. When the polymerase incorporates a fluorescently labeled nucleotide a light impulse is emitted. Light impulses are recorded as a movie.

1.3 Bioinformatics Analysis of Sequencing Data

Conrad Zuse developed the first computer, the Zuse Z1, in 1938. Frederick Sanger and colleagues decoded the first amino acid sequence during 1945 and 1955. The development of high-speed computers by weapons research programs during the Second World War made them highly available for academic research in the 1960s. Bioinformatics history is often stated to have commenced in the 1960s with the “godmother of bioinformatics” Margaret Dayhoff. Dayhoff wrote a program to determine the amino acid sequence of protein molecules in FORTRAN, which computed the correct sequence of ribonuclease within a a matter of several minutes. At this point, computational biologists found a feasible solution to predict protein sequences of up to 750 amino acids in length. Dayhoff established the first sequence database, an annual released version of all known amino acids sequences, reffered to as the “Atlas of Protein Sequence and Structure”, which in 1983 evolved into the online database “Protein Information Resource”. Pairwise comparisons were introduced in the 60s as well as the concept of sequence homology, and alignments of related sequences. Based on these tools and concepts, phylogenetic analyses were born. Walter Fitch implemented the first sliding window approach to accelerate sequence alignments, later improved by Saul Needleman and Christian Wunsch (see section 1.3.1). Based on the Atlas of Protein Sequence and Structure, the “percent accepted mutation” (PAM) matrices were developed, which are still in use [53].

Currently, raw sequencing data, assemblies, structural and functional annotation and taxonimic information are synchronized by the International Nucleotide Sequence Database Collaboration (INSDC), including the DNA Data Bank of Japan (DDBJ at the National Institute for Genetics in Mishima, Japan), the European Nucleotide Archive at the EMBL European Bioinformatics Institute (EMBL-EBI) in Cambridge, UK, and GenBank at the National Center for Biotechnology Information (NCBI) in Bethesda, MD, USA, every day [54]. In October 2018, the NCBI GenBank contained 209,656,636 sequences and 722,438,528 whole genome shotgun sequences and submissions. This reflects the growth of the field of genomics and bioinformatics over the last decades. Biology is one of the fastest growing fields in big data analysis. Electronic engineer Gordon Moore stated that the decrease of costs for components of integrated circuits follows a linear function [55]. This model, also known as Moore's law, is still valid. The complexity of semiconductor microchips doubles every two years for a constant value, while between 2004 and 2010, sequencing capabilities doubled every five months [52]. Since 2008 it is more expensive to store, process and analyze sequencing data than to generate new data. New algorithmic approaches, data compression solutions and parallelization of processes are needed to overcome the challenges of big data processing. The US Government announced a US$200M investment to "improve the tools and techniques needed to access, organize, and glean discoveries from huge volumes of digital data" under President Obama [56]. These trends demonstrate that big data analysis is a challenge for science, society, and politics.

In this study, I focus on sequence alignments, similarity-based sequence clustering, and sequence processing. A sequence alignment is a comparison of two or more nucleotide or amino acid sequences to identify a series of similar characters or a pattern. A selection of sequence alignment approaches will be discussed in the following sections. Alignments are classified as pairwise versus multiple, and global versus local alignments, utilizing optimal versus heuristic algorithms [57].

1.3.1 Pairwise Sequence Alignment

As mentioned before, sequence alignments can be based on optimal and heuristic algorithms. Optimal algorithms will compute the best solution for the string-matching problem. The runtime for sequence alignments based on dynamic programming is quadratic $O(m * n)$, where n is the length of one string and m is the length of the other string. A famous algorithm to compute optimal global alignments was presented in 1970 by Saul B. Needleman and Christian Wunsch [58]. Global alignments compute the best matching string configuration over the complete length of two sequences. The Needleman-Wunsch algorithm can be described as follows:

```
1. Initialization
Create a matrix with M + 1 columns and N + 1 rows where M and N correspond
   to the size of the sequences to be aligned
2. Matrix  fill  (scoring)
Recursion
3. Traceback (alignment)
The maximum score determines the best alignment(s)
```

The Needleman-Wunsch algorithm applies to closely related sequences of similar length. For my work, I used the Smith-Waterman algorithm in chapter 2 to compute optimal local alignments. Local alignments are designed to identify the best match of two substrings. Smith and Waterman first described the Needleman-Wunsch algorithm formally and made two major changes to adjust the algorithm for local alignments [59]. The first column and row of the scoring matrix is initialized with zeros as to not penalize terminal gaps. In addition to the scoring system with match, mismatch and gap, Smith and Waterman introduced a fourth state termed empty suffix.

```
SMITH—WATERMAN OPTIMAL LOCAL SEQUENCE ALIGNMENT pseudocode

Input: two sequences X and Y
Output:optimal local alignment and score $\alpha$
Initialization:
Set F(i,0) := 0 for all i = 0,1,2,...,n
Set F(0,j) := 0 for all j = 1,2,...,m
For i = 1,2,...,n do:
For j = 1,2,...,m do:

$SetF(i,j) := max \begin{cases} 0 \\ F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases}$

Set backtrace T(i,j) to the maximizing pair (i',j')
Set $(i,j) :=$ arg max$\{F(i,j) | i = 1,2,...,n, j = 1,2,...,m\}$
The best score is $\alpha := F(i,j)$
repeat
  if $T(i,j) = (i-1, j-1)$
    print $\binom{x_{i-1}}{y_{j-1}}$
  else if $T(i,j) = (i-1, j)$
    print $\binom{x_{i-1}}{-}$
  else
    print $\binom{-}{y_{j-1}}$
  Set $(i,j) := T(i,j)$
until $F(i,j) = 0$
```

One tool for optimal local sequence alignment is SSearch, in which the Smith-Waterman algorithm is implemented [60]. SSearch was conducted by Pearson who also developed faster, heuristic tools for sequence comparison such as "fasta" and a tool for optimal global sequence alignment termed GGsearch [61, 62, 63, 64, 65].

As indicated, optimal alignments are not feasible for very long sequences or alignments of many sequences. As an alternative, heuristics were developed that do not guarantee an optimal result but are considerably faster. K-tuple alignment methods search for words of size k instead of comparing every character of a string. The underlying logic is to create a database of indexes and subsequences of word size k and to search for a so-called seed (exact match of a subsequence) in the aligned sequences. Smaller k lengths lead to higher accuracy, however, they produce more words which results in slower performance. Seed matches are extended until the accumulated score falls below a specified threshold. Different scoring matrices could be applied, based on assumptions of sequence similarity. In addition to the PAM matrices, the second set of widely used matrices are the “blocks substitution matrices” (BLOSUM) [66]. While PAM is based on global alignments, BLOSUM is based on local alignments and therefore applicable to more distantly related sequences. One of the most prominent heuristic alignment software is the basic local alignment search tool (BLAST) released 1990 by Stephen Frank Altschul [67]. Different features have been implemented since then, such as parallel mode, GPU mode and different modules for protein alignments, nucleotide alignments and cross-comparison [68, 69, 70]. BLAST remains the gold standard for a broad group of users, even though there is more efficient sequence alignment software, such as diamond [71] or NSimScan [72].

Since it is not possible to compute an exact value for the precision of an alignment without knowing the optimal solution, probability scores are computed for every heuristic. It is crucial to evaluate the quality of a non-optimal alignment. The alignment quality measure of BLAST is the so-called e-value, short for “expect value”. The e-value is the number of expected hits of similar quality. It is calculated by a weighted ratio of the bit-score, the number, and length of sequences in the database and the query sequence length. The bit-score ($S' = \frac{\lambda S - lnK}{ln2}$, where K and λ are parameters calculated from scoring matrix) is a log2 scaled, normalized raw-score (S), independent of the database size and describes the probability of the current alignment to occur by chance [67]. Given that m is the total length of sequences in the database and n is the query length, the e-value is calculated according to the formula:

$$e = \frac{m * n}{2^{-S'}} \tag{1.1}$$

E-values are used to compare alignments with each other, independent from database size.

Even though BLAST has the same overall computational complexity of $O(m * n)$, in practice, BLAST is faster by orders of magnitudes than dynamic programming approaches. The development of more efficient algorithms such as SeqAn libraries may boost the performance of alignment software, leading to more precise and faster tools [73, 74].

1.3.2 Multiple Sequence Alignment

In contrast to pairwise alignments, a multiple sequence alignment applies to a set of more than two sequences compared with each other and thereby producing a set of alignments with the smallest distances among all pairs of sequences. In 1994, it was shown that multiple sequence alignment with sum of pairs score is NP-complete and multiple tree alignment is MAX SNP-hard [75, 76]. This indicates that an alignment with a sum of pair score (SP-score) cannot be solved in polynomial time, whereas multiple tree alignments allow polynomial-time approximation. Due to this fact, heuristics are needed to solve multiple sequence alignments, in contrast to pairwise sequence alignments for which optimal alignments are still feasible.

Progressive and consistency-based tools can be pooled as similarity-based approaches and are most widely used. MAFFT [77] and clustal [78, 79] are examples of progressive multiple sequence alignment tools, whereas T-Coffee [80, 81] is a consistency-based alignment tool. Consistency-based alignments compute pairwise alignments; i.e., based on the alignments of sequence A to B and B to C can the resulting distance of sequence A to C be justified by computing the distance by aligning sequence A to C. Progressive alignments compute pairwise alignments of the most similar pairs of sequences. Creating a so-called guide tree, alignments of interior nodes with direct descendants are computed. The root node represents a complete multiple sequence alignment. This can be refined by reiterating this process based on the last computed guide tree.

Iterative refinement concepts, e.g. MUSCLE [82], are another class of multiple sequence alignment strategies. In contrast to progressive alignments initially computed pairwise alignments can be split again, realigning a sequence if the distance to another sequence or group of sequences is smaller than the initial alignment. As a result, a local optimal score is produced, emerging into a maximum global space as sequence space is finite. MAFFT and MUSCLE offer significant improvements in scalability with comparable accuracy and thus provide reasonable starting points for general alignment problems.

The score of an alignment can be calculated either using an evolutionary model [83] or with the SP-score. Therefore, the Needleman-Wunsch scoring matrix as described in section 1.3.1 was adapted. Historically, most algorithms assumed that sequences align globally. Nowadays, algorithms exist which focus on sequences with local similarities as well, including ALIGN-M, DIALIGN, POA or SATCHMO; this, however, is beyond the scope of this thesis [84].

Recent approaches such as PRANK [85] focus on phylogeny-aware alignments. Phylogeny-based and probabilistic approaches tend to be the next level of multiple sequence alignments and can be joined as evolution-based alignments [86]. For phylogenetic based approaches, a set of N sequences is aligned by performing N-1 pairwise alignments computing a phylogenetic tree connecting the sequences. Iterative refinement is computed until the phylogenetic tree reaches a steady state. In contrast, probabilistic models implemented in e.g. HMMER [87] extend entropy-based scores by modeling insertions and deletions in multiple sequence alignments with hidden Markov Models, estimating a position-specific model weighted with probability scores.

Specialized multiple sequence alignment tools such as MARS [88] were developed for specific problems, which cannot be addressed using standard multiple sequence alignment tools. MARS is a heuristic method for improving multiple circular sequence alignment using refined sequences. It was developed to relax the assumption that the start position of the alignment is at the first position of the sequence and the alignment end is at the end of the sequence. This assumption does not necessarily apply to circular sequences, and therefore MARS computes the cyclic edit distance between two strings and finds the best rotation of the sequences to minimize pairwise distance.

1.3.3 Sequence Mapping

Sequence mapping is the process of aligning numerous queries to a small amount of longer sequences. Classical sequence mappers have been adapted from the needs of short reads to long reads within the past years. A commonly used tool for sequence mapping is Bowtie [89]. Bowtie creates its own refined index based on the FM Index, which uses the Burrows Wheeler transformation. Bowtie1 was designed for short reads of up to

1 kbp length, whereas Bowtie2 has no upper limit regarding read length. For my analyses Burrows-Wheeler Aligner's Smith-Waterman Alignment (BWA-SW) [90] was used to align long sequences against references. Burrows–Wheeler transform is the first step to reversibly compresses strings, thereby reducing search space and consequently reducing memory usage in order to accelerate the process. BWA produces sequence/alignment maps (SAM) [91] containing information on start- and end-position of local alignments, global and local quality information of alignments and could be run in paired mode, such that distance information of sequences from paired end libraries can be facilitated. A different long-read mapping tool is BLASR (basic local alignment with successive refinement) [92]. BLASR is specialized on long reads from HTS of DNA by single molecule sequencing. In 2018, a tool termed Minimap2 was released which is more than 50 times faster and more accurate for long reads than bwa-mem [93].

1.3.4 Sequence Clustering

Sequence clustering attempts to group biological sequences by similarity. Uclust [94] and cd-hit [95] use greedy algorithms to identify representative sequences and assign related sequences based on a given threshold of identity. Tribe-MCL [96] relies on the Markov Cluster (MCL) algorithm ,[97]. MCL is able to utilize BLAST scores, such that all versus all database searches using BLAST are used to compute reliable sequence clusters. In brief, Markov Models or Markov matrices describe the transitions of Markov Chains [98]. A Markov Chain makes predictions based exclusively on the present state, such that future and past states are independent. Also, Hidden Markov Models (HMM) include unobserved (hidden) states. It has been shown that HMMs represent the evolution of biological sequences very well [83]. HMMs are able to predict state changes, based on the sequence of observations, whereas internal hidden stages predict a set of external events (observations). A small example of an HMM for a DNA sequence is shown in figure 1.3. A set of probabilities for every nucleotide at a certain position is given (observation/emission probabilities - match state), as well as probabilities for insertions (insert state) and deletions (transitions - delete state).

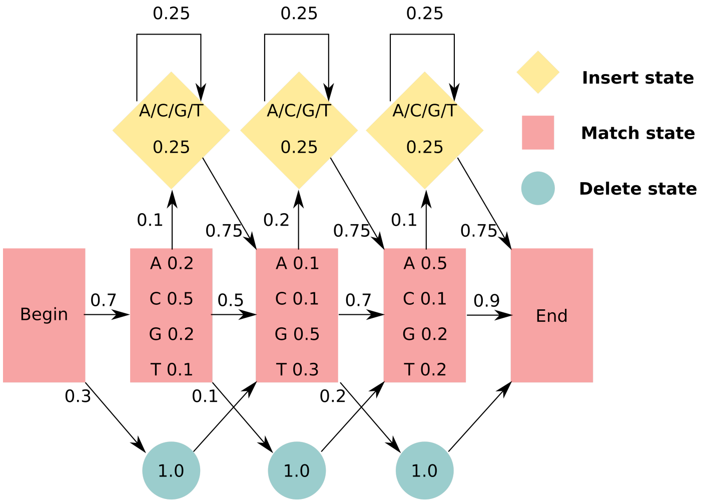


Figure 1.3: Schematic Hidden Markov Model

A simplified Hidden Markov Model showing probabilities of residue types for individual positions (match state red rectangles, insertion state yellow rhombus, deletion state green circles).

1.3.5 Tools and Pipelines for Retrieval of Retroviral Integration Sites

Different approaches can be taken to retrieve viral integration sites. SeqMap is a publicly available web platform for retroviral integration site analysis of HTS data using ligase-mediated PCR (LM-PCR) [99], linear-amplification-mediated PCR (LAM-PCR) [100, 101] or non-restrictive LAM-PCR (nrLAM-PCR) [102]. All of these amplification protocols depend on the presence of specific recognition motifs that are unevenly distributed across the genome. SeqMap is a reference-based integration site mapper [103]. Basically, it performs sequence preprocessing including demultiplexing, recognizes LTR sequences and linker cassettes by sequence alignment, masks adapters, LTRs, and linker cassettes and maps these reads to a reference genome. Results are visualized and stored for further analysis. In the latest version, SeqMap 3.0, “mouse” (mm10) and “human” (hg19) are the only implemented references, however, other references might be implemented on request. This approach is neither applicable for non-model organisms, nor for fragmented DNA or third-generation sequencing technologies. Another tool to identify retroviral sequences in host genomes is RetroTector [104]. However, RetroTector is limited to nine predefined reference species and can handle a maximum of 100 Mb of data per run. VISPA (Vector Integration Site Parallel Analysis) is a pipeline to identify genomic vector integration sites [105]. The pipeline was implemented to handle reads of 100 to 1,000 bp length produced by LAM-PCR [100, 101].

There are several other tools to detect retroviral integration sites, such as VirusSeq [106], ViralFusionSeq [107] VirusFinder [108], Virus-Clip [109], Ub-ISAP [110] and ViFi[111]. All of these are designed to identify viruses in the human genome, thus they are reference-based and mostly restricted to RNA sequencing to detect proviruses of exogenous retroviral origin. Other approaches designed for wild and domestic animals are at least confined to an assembled reference genome [112].

To the best of my knowledge, no tool exists to detect endogenous retroviruses in large scale high throughput data from ancient DNA or long-read sequencing for non-model organisms.

1.4 Objective

This thesis aims to investigate further methods to process data from targeted short and long-read HTS to compare insertion sites of endogenous retroviruses comprehensively. By doing so, I aimed to address the underlying mechanisms of retroviral endogenization in the earliest stages of genomic invasion. In order to achieve this objective, I developed computational pipelines to detect endogenous retrovirus integrations from HTS data. A pipeline for short-read sequences from historical koala samples is described in chapter 2. A new methodology for target enrichment sequencing using third-generation sequencing technology is outlined in chapter 3. The following chapter (chapter 4) examines aspects of retroviral recombination and compares findings retrieved from data generated by the method described in chapter 3. My conclusions are discussed in the final chapter (chapter 5).

1.4.1 Retroviral Integration Sites

Research tended to focus on the development and comparison of target enrichment techniques of endogenous retroviruses in modern samples rather than in historical samples. Moreover, it remains unclear whether standard methods like rapid amplification of cDNA ends (RACE) [113], ligation-mediated PCR [114], linker-selection-mediated PCR [115], linear amplification-mediated PCR [101] and genome walking [116] comprehensively detect integration sites, because of the potential of primer-target mismatches [117]. In order to investigate other methods for retroviral target enrichment from ancient DNA, we tested three different target enrichment methods, namely primer extension capture, single primer extension capture, and hybrid capture, in order to comprehensively profile integration sites of the koala retrovirus from museum samples. A severe limitation of analyzing ancient DNA is the fragmentation of the DNA resulting from natural degradation processes. As one outcome the target enrichment techniques are compared, and as another outcome, koala retrovirus integration sites were profiled, by combining novel and published data.

Since new approaches in sequencing technologies have been developed, it is now possible to produce long sequences with high-throughput technologies. The main shortcoming of short-read HTS, particularly the length limitation of sequences thus could be resolved. New methods need to be developed to overcome the constraints of target enrichment methods for short-read sequencing. I propose a new technique to capture retroviral integration sites in chapter 3. Sonication inverse PCR is a restriction enzyme free approach for the genome scale analysis of integration sites using long-read sequencing technologies. I describe the method and the bioinformatics pipeline I developed and discuss limitations. In chapter 4 I outline findings retrieved by sonication inverse PCR applied to the genome of a captive koala. Furthermore, the results are compared with koala retrovirus integration sites from literature. I also describe the discovery a new retroviral recombinant.

Problems and remaining questions regarding the processing of retroviral integration sites from ancient DNA as well as sequences produced by sonication inverse PCR are discussed in chapter 5. Research perspectives and open questions are summarized in a concluding section at the end of this thesis.

Chapter 2

Investigation of Koala Retrovirus in Museum Koala Samples

For hundreds of millions of years, endogenous retroviruses (ERVs) have been invading vertebrate genomes [4]. In contrast to exogenous retroviruses, ERVs infect germline cells and are inherited as a Mendelian trait. The koala retrovirus (KoRV) is currently in transition from an exogenous to an endogenous retrovirus [118, 119]. It has been shown that KoRV causes neoplasia[120], lymphoma and leukemia [27], increases the prevalence of chlamydia infections and leads to immunomodulation [28, 29]. In recent years there has been considerable interest in KoRV [31, 121, 122, 123, 124]. Previous work has only focused on the examination of KoRV integration sites in different koalas but failed to address a comprehensive characterization between individuals. Different methods for target enrichment sequencing exist but have not been applied previously to ancient DNA. In order to close these gaps, this work aims to evaluate different methods of targeted high-throughput sequencing (HTS) from historical samples to comprehensively profile KoRV with new and published data. Comprehensive profiling is necessary to shed light on the distribution of unique KoRV integration sites and those shared between individuals.

Paleogenomics analyses are challenging, due to the fact that ancient DNA is typically heavily fragmented by endogenous nucleases, oxidation, hydrolysis and background radiation [125]. As a consequence, computational solutions must be found to process sequences and assign integration sites from very short reads. A major limitation is the contamination of target DNA, such that extensive preprocessing of the molecular data is necessary.

2.1 Characterizing Viral Integration Sites from Ancient DNA

This chapter is based on the publication “Comprehensive profiling of retroviral integration sites using target enrichment methods from historical koala samples without an assembled reference genome.” published in the journal PeerJ in 2016 [1]. In recent publications it has been shown that KoRV integration sites were only shared in very closely related koalas. Nonetheless, no comprehensive integration site analysis was performed so far, since methodological challenges limited the analysis space. At this point, there was no reference genome available for the koala. Target enrichment techniques for retroviral integration sites have been tested, including inverse PCR [126], and methods such as

Table 2.1: Museum Koala Sample Information [1]

Collection no.	Year	Sample provider	Locality
AMA17300	1883	Australian Museum	New South Wales, Australia
AMA17311	1883	Australian Museum	New South Wales, Australia
AMA17299	1883	Australian Museum	New South Wales, Australia
QM J2377	1915	Queensland Museum	Queensland Australia
QM J7209	1945	Queensland Museum	Queensland Australia
QM J8353	1952	Queensland Museum	Queensland Australia
QM JM1875	1960s	Queensland Museum	Queensland Australia
AM M 12482	1971	Australian Museum	New South Wales, Australia
QM JM64	1973	Queensland Museum	Queensland Australia
QM 7625	1970–1980s	Queensland Museum	Queensland Australia

rapid amplification of cDNA ends, ligation-mediated PCR [114, 127], linker-selection-mediated PCR [115], linear amplification-mediated PCR [128], and genome walking [129]; however, none of these were applied to ancient DNA (aDNA) so far. To investigate KoRV integration site patterns we examined ten museum koala samples (s. table 2.1). We developed and tested three different target enrichment techniques for HTS of short reads. In this thesis, the focus is on the development of the bioinformatics pipeline to analyze short-read products from ancient ERVs.

2.1.1 Comparison of Three Targeted Sequencing Methods

Three target enrichment techniques followed by Illumina HTS were applied - Single Primer Extension (SPEX) [130], Primer Extension Capture (PEC) [131] and Hybridization Capture (HC) [132]. All of these techniques have successfully been applied to enrich aDNA and were tested for the characterization of genomic sequences flanking ERVs. Museum skin samples of ten koalas were examined for this study. Our primary aim was a cross-comparison of the efficiency of target enrichment techniques. Furthermore, we retrieved integration site information on historical koala samples, and compared them with each other and to known integration sites of modern koalas of other studies. All enrichment techniques were based on DNA from a 7 mm × 7 mm sample of museum koala skin. Primers and baits were designed to bind to the LTR region of KoRV. Sequencing was performed using an Illumina MiSeq Reagent Kit v2 for all experiments. Underlying mechanisms are described in the following sections and visualized in figure 2.1.

Primer Extension Capture (PEC)

For Primer Extension Capture, Illumina libraries are produced from DNA and are subsequently denaturated to produce single-strand DNA. Sequences containing the target region hybridize with biotinylated primers which are then bound to magnetic beats so that non-targets can be removed by washing. The target product is eluted and sequenced.

Single Primer Extension Capture (SPEX)

In contrast to PEC, the Illumina library preparation is the last procedure of the protocol before SPEX products are created. DNA is denaturated and mixed with the same biotinylated primers as for PEC. Targets are filtered using magnetic beats, and a poly C tail is ligated. The Illumina library is constructed, and a second amplification is performed. Then the target products are sequenced.

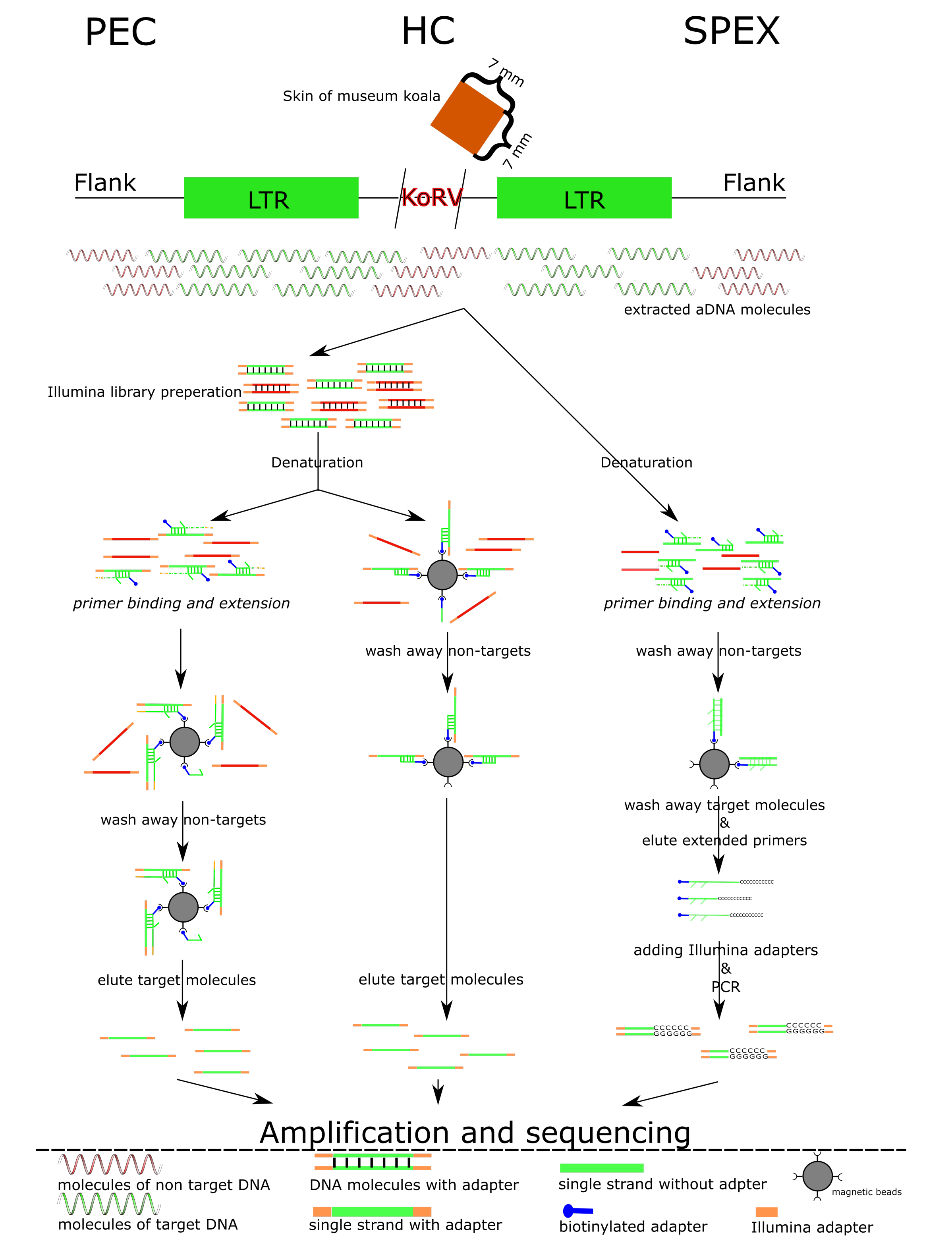


Figure 2.1: Experimental Workflow for Three Target Enrichment Techniques [1]

Table 2.2: Sequencing Statistics of Target Enrichment Techniques[1]

	SPEX	PEC	HC
Number of raw reads	7,627,810	6,956,280	31,096,064
Unique sequences	714,929	1,188,365	11,675,245
Target enrichment efficiency (%)	4.68	0.55	0.01
Homologous sequences to wallaby genome	1,617	136,366	1,915,781

Hybrid Capture (HC)

As for PEC, the first step of HC is the construction of an Illumina library. After denaturation, single-stranded DNA is mixed with rotating magnetic beats with immobilized baits on their surface. The baits bind target sequences and a hybridization process starts where captured baits bind further downstream target sequences. After hybridization, the products are eluted and sequenced. In contrast to PEC and SPEX, this process leads to a bell-shaped distribution of target sequences adjacent to the bait.

2.1.2 Detection of the Koala Retrovirus from Illumina Short-Reads

The focus of this thesis is the analysis regarding approaches of applied bioinformatics for automated detection of ERV integration sites within mammalian genomes exemplary demonstrated for the case of KoRV. I developed a pipeline to detect ERVs from ancient koala samples using different target enrichment techniques and Illumina HTS. To the best of my knowledge, no standard protocol was published to tackle this problem. Existing approaches for retroviral detection such as RetroTector have several limitations. RetroTector online can process "up to 100 Mbase per submission" which is not applicable to HTS data [104, 133]. RetroTector is based on the identification of both LTRs assigning domains in-between, which requires complete, and intact proviral structures. Other approaches are alignment-based using viral references, which are designed for complete genomes, as well or at least assembled contigs [17]. However, aDNA is fragmented, such that reference-free assemblies, especially for repetitive or low complexity regions, would fail.

A total of 7,627,810 reads from SPEX, 6,956,280 reads from PEC and 31,096,064 reads from HC respectively were analyzed (s. table 2.2). aDNA was sequenced using paired-end libraries with Illumina MiSeq Reagent Kit v2 producing sequences up to 150 bp in length. The sequence length distribution of all three enrichment techniques is shown in figure 2.4.

All samples were processed separately until the final comparison steps of KoRV integration sites. Cutadapt (version 1.2.1) [134] was applied for adapter trimming; quality filtering was performed using Trimmomatic (version 0.22) [135], both with default settings. Forward and reverse reads were merged using Flash (version 1.2.5) [136]. Downsampling was performed using cd-hit (version 4.6) [95] to screen for identical sequences, retaining the longest representative read. Downsampling enhances the performance of follow up processes, due to the smaller number of operations, as the objective is a qualitative and not quantitative analysis.

After preprocessing, off-target reads were filtered, and KoRV homologous regions were clipped. The reads were aligned to a 30 bp reference on the 5' site and 63 bp reference on the 3’ site extracted from the LTR covering the bait binding sites. A pairwise alignment using the Smith-Waterman [59] application from EMBOSS (version 6.6.0.0) [137] was executed. The results were further filtered based on different criteria for 5’ and 3’ integration sites as the baits bind 49 bp (apart from the 5’ breakpoint) and 82 bp (apart from the 3’ breakpoint), respectively. Primer positions relative to the LTR structure of KoRV are shown in figure 2.2.

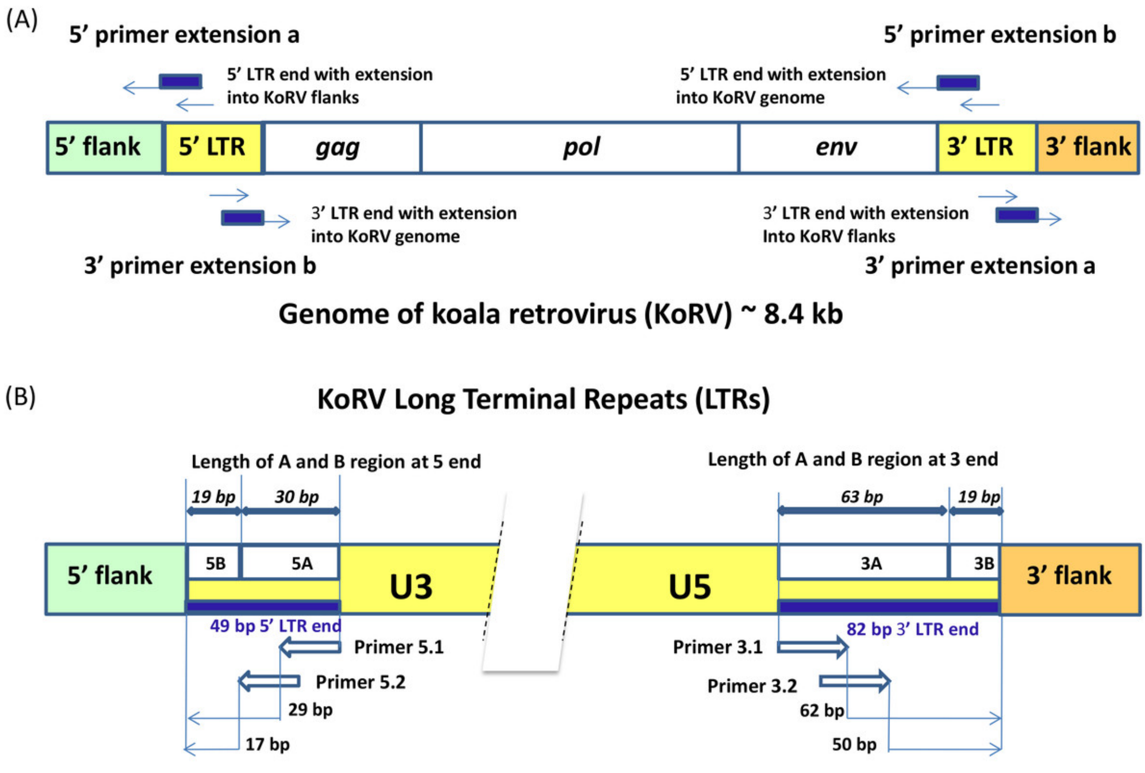

(A)
5’ primer extension a
5’ primer extension b
5’ LTR end with extension into KoRV flanks
5’ LTR end with extension into KoRV genome
5’ flank
5’ LTR
gag
pol
env
3’ LTR
3’ flank
3’ LTR end with extension into KoRV genome
3’ LTR end with extension Into KoRV flanks
3’ primer extension b
3’ primer extension a
Genome of koala retrovirus (KoRV) ~ 8.4 kb
(B)
KoRV Long Terminal Repeats (LTRs)
Length of A and B region at 5 end
19 bp
30 bp
Length of A and B region at 3 end
63 bp
19 bp
5’ flank
5B
5A
U3
U5
3A
3B
3’ flank
49 bp 5’ LTR end
82 bp 3’ LTR end
Primer 5.1
Primer 5.2
29 bp
17 bp
Primer 3.1
Primer 3.2
62 bp
50 bp

Figure 2.2: Primer Positions and LTR Structure for the Identification of KoRV Integration Sites [1]

(A) General overview of KoRV structure and primer position overview (B) Detailed primer positions with respect to gDNA and LTR

It was shown that some KoRV proviruses had a 19 bp deletion in the LTR towards the breakpoint [31]; we therefore, evaluated filtering criteria based on these factors and filtered reads which covered at least two thirds of the LTR reference (20/30 bp on the 5’ site and 43/63 bp on the 3’ site) with 90 percent identity. Sequences not matching these criteria were discarded and identified LTR domains were clipped. The remaining sequences were aligned to the residual 19 bp of the LTR sequence clipping sequences of 12/19 bp could be aligned with a minimum identity of 80 percent. All routines were implemented in Perl. An overview of the pipeline is shown in figure 2.3.

Table 2.3: megaBLAST and TRIBE-MCL Parameters for Sequence Clustering [1]

Method	SPEX		PEC		HC	
Integration site orientation	5' end	3' end	5' end	3' end	5' end	3' end
E-value for all versus all BLAST	10^{-30}	10^{-30}	10^{-17}	10^{-20}	10^{-15}	10^{-15}
Inflation value for clustering	1.4	4	22	4	6	16

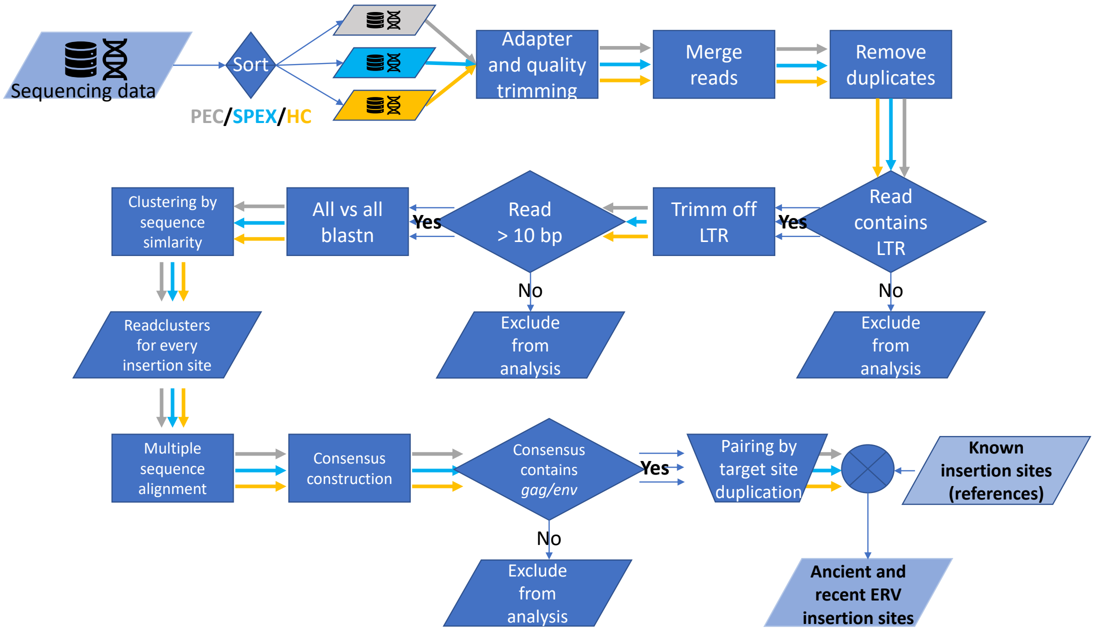

Sequencing data
Sort
PEC/SPEX/HC
Adapter and quality trimming
Merge reads
Remove duplicates
Read contains LTR
Yes
No
Exclude from analysis
Trimm off LTR
Read > 10 bp
Yes
No
Exclude from analysis
All vs all blastn
Clustering by sequence simlarity
Readclusters for every insertion site
Multiple sequence alignment
Consensus construction
Consensus contains gag/env
Yes
No
Exclude from analysis
Pairing by target site duplication
Known insertion sites (references)
Ancient and recent ERV insertion sites

Figure 2.3: Bioinformatics Pipeline for Targeted ERVs from Museum Samples

Starting with raw sequencing data, all steps were performed separately for PEC, SPEX and HC, indicated by the yellow, blue and grey arrows, respectively. Rectangles indicate processes, the rhombus indicates decisions (filtering), the parallelogram indicates data, the trapezium indicates manual processing and the circle indicates coalescence of new and published data.

Reads longer than ten base pairs were used for further analyses. We created a distance matrix for the sequences with TRIBE-MCL (version 12-135) [96] based on NCBI megablast alignments (megaBLAST from NCBI BLAST+ version 2.2.29+) [67, 68, 138, 69] using all reads as database and query for the three different datasets.

For each resulting cluster, a multiple sequence alignment (MSA) was performed using MAFFT (version 7.127b) [77, 139] with default settings except for adjusting for reverse complement sequences. The clustering parameters were assessed by visualizing the MSAs of the 30 largest clusters in jalview (version 2.8) [140]. When no sequence in each of these clusters had less than 10 percent insertions, deletions or substitutions related to the individual sequence length, the parameters with the lowest granularity matching these conditions were chosen. MSAs were further processed to construct consensus sequences for every cluster . Clusters containing only one sequence were considered “singletons”. Consensus sequences and singletons, were aligned with the sequences of group-specific antigen gene (*gag*) and envelope gene (*env*). If one of the genes could be assigned to a consensus sequence or singleton, the cluster was assigned as proviral primer extension and was excluded from the analysis. Due to the duplication of the LTR and the primer design we expected approximately 50 percent of the products from proviral origin while the other products are from genomic DNA (gDNA) integration sites, respectively. Further, non-automated processing and statistical analyses were performed as described in Cui et al. 2016 [1].

2.2 Results

This study had two objectives. Most studies have only focused on target enrichment techniques to find ERV integration sites in fresh samples. The first objective was to compare three target enrichment techniques to identify ERV integration sites in museum koala samples. The characteristics of ERV integration sites have not been dealt with in depth. The second objective is the comprehensive profiling of KoRV integration sites from different koalas. Both objectives were achieved.

2.2.1 Target Enrichment Techniques for Integration Site Retrieval from Ancient DNA

PEC and HC experiments showed similar sequence length distributions with the most frequent fragment length close to 50 bp and 150 bp, respectively. The maximum sequence length produced by the Illumina MiSeq Reagent Kit v2 was 150 bp. Sequences from the SPEX experiment showed a slightly different sequence length distribution with most reads of 20 bp in length. Similar to PEC and HC a higher frequency of reads with 50 bp and 150 bp in length occured. Additionally, a high abundance of reads with a fragment length of 90 bp were retrieved from SPEX. Reads with a sequence length of 150 bp were observed relatively more frequently in SPEX than in PEC or HC. All observations are shown in figure 2.4 and were normalized to the total number of reads in every dataset.

The highest amount of sequences was generated with the HC protocol (31,096,064 sequences), followed by SPEX (7,627,810 sequences) and PEC (6,956,280 sequences). After preprocessing (adapter and quality trimming, read merging and dereplication) 11,675,245 (38%) of the reads from HC, 7,627,810 (9%) from SPEX and 1,188,365 (17%) from PEC were retained. These numbers indicate that SPEX produced the highest amount of clonality, whereas HC produced both the biggest number of reads and the smallest amount of clonality, which was expected due to methodological assumptions (see figure 2.5).

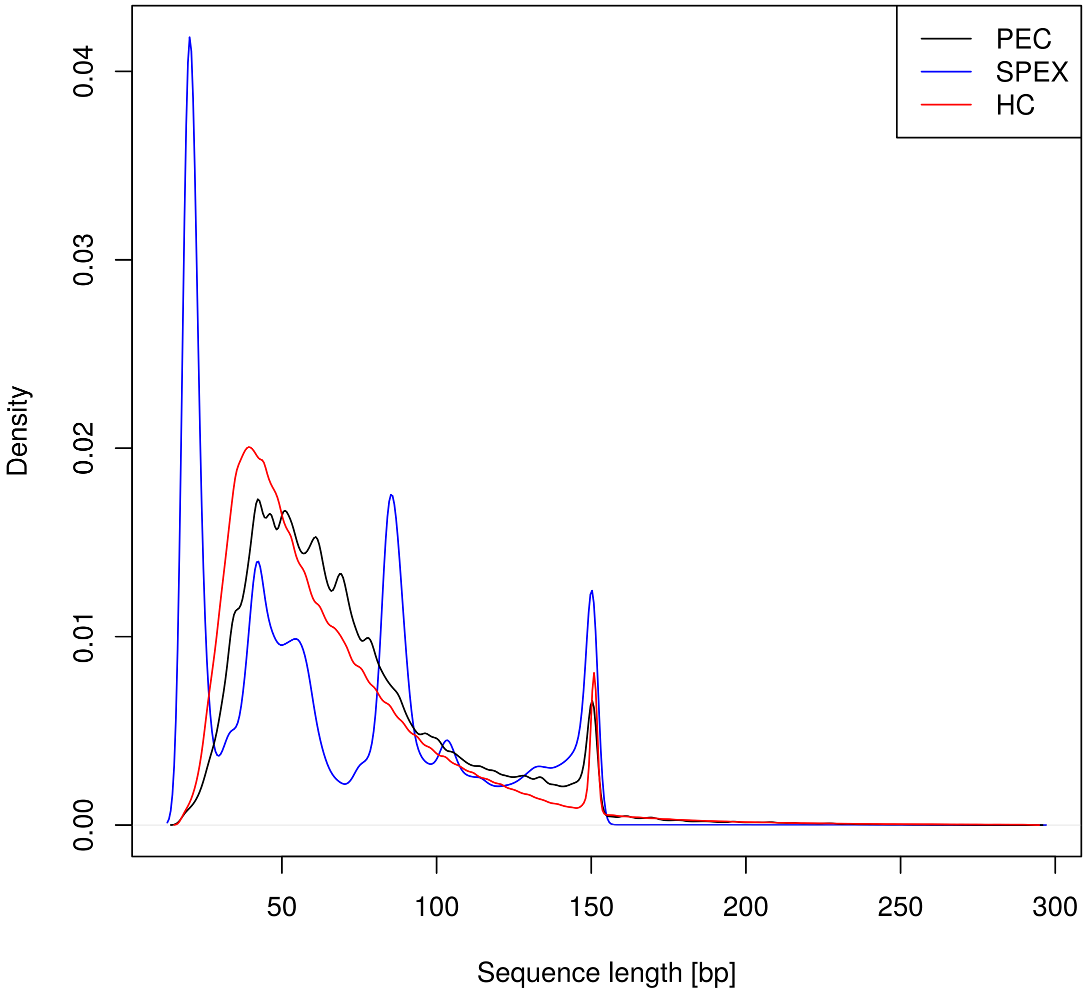


Figure 2.4: Sequence Length Distribution of Three Target Enrichment Techniques [1]

Sequence length distributions of reads from PEC (black), SPEX (blue) and HC (red). Sequences from the SPEX experiment show a slightly different sequence length distribution with a peak of short sequences around 20 bp length which are at low frequency in PEC and HC. All techniques amplified reads of approximately 50 bp in length. A third sequence length with high abundance is 90 bp for SPEX. Reads with a sequence length of 150 bp are observed more often in SPEX than in PEC or HC.

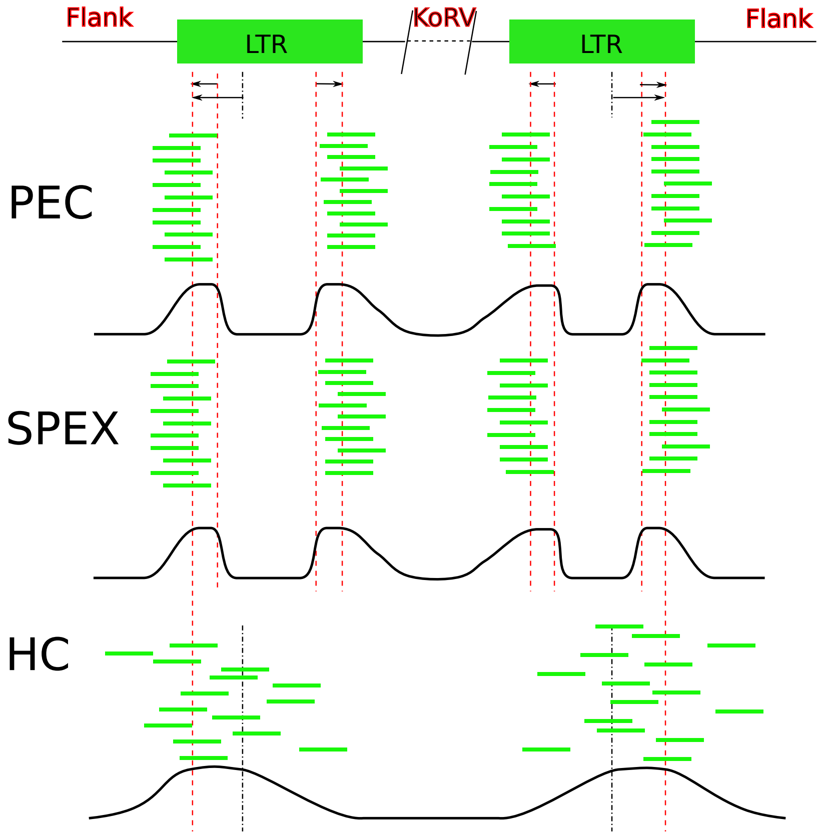


Figure 2.5: Expected Product Distribution of Target Enrichment Techniques

The oligos used for all three experiments bind near to the end region of the KoRV LTR. Arrows indicate primer binding sites. Due to the duplication of LTRs on both ends, primer extension of captured products using these oligos will result in products from koala DNA flanking KoRV and internal KoRV reads. The bold black line at the bottom of each section presents the expected distribution of target sequences.

In each experiment, we observed a bias towards one end of the integration sites; for SPEX and PEC far more unique 3' integration sites were detected than 5' integration sites, while HC spotted more unique 5' integration sites than 3' integration sites. In contrast to the expected product distribution shown in figure 2.5, table 2.4 lists the number of retrieved 3' and 5' integration sites.

Integration site sequences shorter than four base pairs were too short for any further biological interpretation, and they accounted for 52 percent in SPEX, 25 percent in PEC and 15 percent in HC. Sequences of a minimum length of four base pairs were used to calculate the number of shared and unique KoRV integrations across individuals, but only integration site sequences of 15 bp or longer (399 sequences SPEX/377 sequences PEC/110 sequences HC) were used for pairing based on target site duplications.

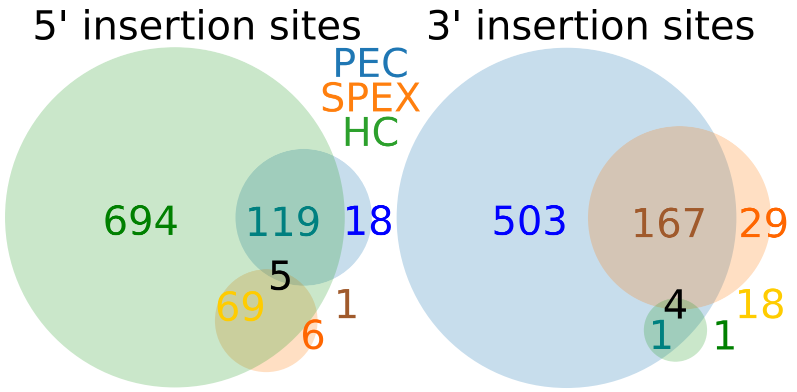


Figure 2.6: Overview of Integration Sites from Ten Individuals Across Target Enrichment Techniques

Left Venn diagram shows 5' integration sites. HC (green) retrieved the highest amount of integration sites, and covered 91 percent of the integration sites found by SPEX (orange) and 87 percent of the integration sites found by PEC (blue). Overlapping circles and mixed colors indicate intersections. Intersection for 5' integration sites PEC/SPEX is only represented by the count (1). Intersection for 3' integration sites SPEX/HC is only represented by the number (18). Right Venn diagram shows 3' integration sites retrieved from target enrichment techniques. The most significant amount of integration sites was recovered from PEC, covering 81 percent of integration sites retrieved from SPEX and 91 percent from HC.

Table 2.4: Sequence Statistics for Integration Site Detection from aDNA [1]

	SPEX		**PEC**		**HC**	
KoRV flanks orientation	5' end	3' end	5' end	3' end	5' end	3' end
KoRV flanks <4 bp	15,822	1,527	496	1,806	191	41
KoRV flanks 4–14 bp	6,426	8,896	329	2,033	1,052	24
KoRV flanks >=15 bp	95	304	63	314	106	4
Internal KoRV reads	212	223	141	1,406	151	14
Target enrichment products	22,542	10,950	1,029	5,559	1,495	83
*Clustered unique integration sites	66	182	126	538	862	24
*Clustered shared integration sites	15	28	17	134	25	0

*****_unique/shared integration sites across koala individuals_

A comparison of unique integration sites across target site enrichment techniques is visualized in figure 2.6, referring to table 2.4. Most unique 5' integration sites were detected using HC, while PEC enriched for the highest amount of 3' integration sites. Five unique 5' integration sites were shared between all three target enrichment techniques, while four 3' integration sites were shared among all three target enrichment protocols. Thus, HC covered nearly all produced 5' integration sites, while PEC covered nearly all 3' integration sites.

As a conclusion regarding the comparison of three target enrichment protocols, we recommend a combination of PEC and HC. Most integration sites captured using SPEX could be reproduced by HC regarding 5' integration sites and by PEC for 3' integration sites. 6/912 (1%) of 5' integration sites would have been missed, whereas 29/723 (4%) of the 3' integration sites would have been missed.

2.2.2 Comprehensive Profiling of Koala Retrovirus (KoRV) Integration Sites

Clustering resulted in 43 integration sites shared across individuals for SPEX, 151 shared integration sites across PEC and 25 shared integration sites across HC (see table 2.4 *). Unique integration sites were composed of sequences from one individual or are represented as singletons. In general, there were more unique than shared integration sites across individuals. It was observed that SPEX produced substantially more reads for 5' integration sites and resulted in more 3' than 5' integration sites after clustering, which suggests that SPEX tends to overamplify 5' integration sites.

We computed around 1650 5' and 3' integration sites of which 63 aligned to unassembled koala shotgun sequencing (HiSeq 100x coverage) data from a Queensland koala. Twenty-three of these integration sites were present in more than one koala. In this study, 865 unique and 52 shared 5' integration sites were retrieved across individuals. The rate of 5.7 percent shared integration sites is in accordance with published observations [123]. However, 20.4 percent 3' integration sites were shared. Five hundred seventy unique 3' integration sites and 146 shared 3' integration sites were classified.

Most of the integration sites were shared between two koalas (32 5' integration sites, 80 3' integration sites) as shown in table 2.5. Four 5' and two 3' integration sites were found in every individual tested in this study. The median of individuals per shared integration sites was 1.5 for 5' and 5.5 for 3' integration sites. This indicates that not only the percentage of shared integration sites is higher for the 3' site, but also that the retrieved integration sites were shared by more individuals.

Retrieved 5' and 3' integration sites from this study were compared to historical and modern samples from Tsangaras et al. (2014) [123] and modern samples from Ishida et al. (2015) [141]. Among all compared integration sites, ten integration sites are shared between individuals from our study and individuals analyzed in the aforementioned studies. Two individuals (*QM J7209*, *QM JM64*) share one integration site with a non-historical sample (*Pci-SN265*) from Tsangaras et al.. One of these individuals was *QM J7209*, which was sampled in 1945 and was provided by the Queensland Museum (Australia). *QM J7209* shares another integration site with two historical samples [123], and two other integration sites were shared twice across three modern samples [141]. *QM JM64* shared a second integration site with *Pci-SN265* [123]. Two individuals (*Pci-SN404*, *Pci-SN345*) from Ishida et al. shared two different integration sites with *QM M187* from this study, whereas each of these integration sites was present in one additional individual from our study (*QM J2377*, *QM J8353*).

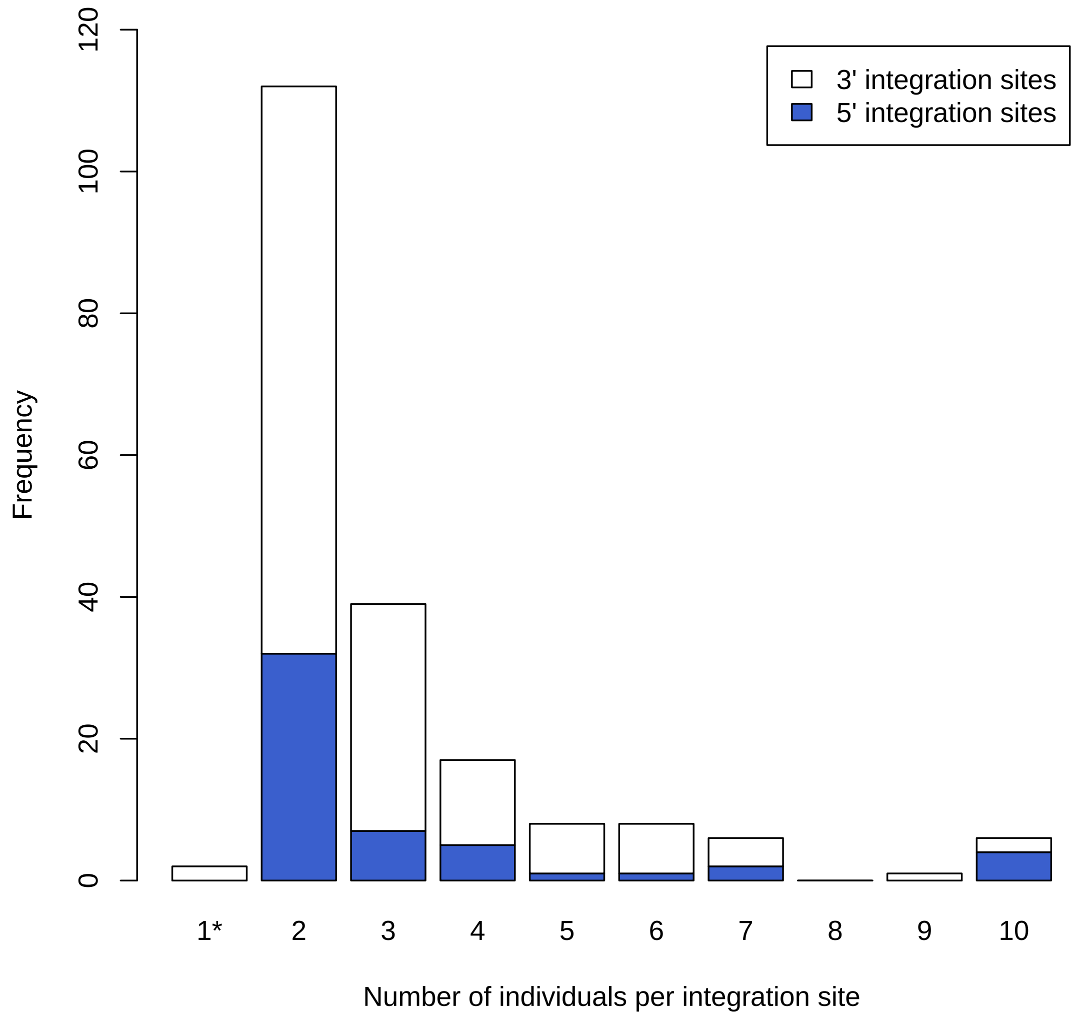


Figure 2.7: Number of Individuals from this Study per Shared Integration Site

**Two 3' integration sites are indicated to be "shared" by only one koala, in fact, these integration sites are shared with koalas from other studies*

Two of the analyzed museum samples share integration sites with two modern samples, reported in Ishida et al. In addition, we identified one integration site shared between two of the modern koalas, one reported in Tsangaras et al. and one reported in Ishida et al. (see table 2.5 and 2.1 for sample information). Additionally, one integration site was shared among two of the samples from this study, three modern samples from Ishida et al., one modern and one historical sample from Tsangaras et al. One integration site was shared among four of our samples, one modern sample reported in Ishida et al. and four historical samples from Tsangaras et al. (not shown in table 2.5).

Table 2.5: Shared Integration Sites Across Studies

QM J7209		QM JM1875		Pci-SN265	[123]		
QM J7209		maex 1738	[123]	582119	[123]		
QM JM64		QM JM1875		Pci-SN265	[123]		
QM JM64		Pci-SN265	[123]				
QM JM1875		QM J2377		Pci-SN404	[141]	Pci-SN345	[141]
QM JM1875		QM J8353		Pci-SN404	[141]	Pci-SN345	[141]
QM J7209		Pci-SN404	[141]	Pci-SN345	[141]		
QM J7209		Pci-SN404	[141]	Pci-SN248	[141]		
Pci-SN265	[123]	Pci-SN248	[141]				
Pci-SN265	[123]	Pci-SN248	[141]	Pci-SN404	[141]		

2.3 Discussion

The processing of short-reads for retrieval of endogenous retroviruses is generally limited due to the occurrence of LTRs on both ends of the provirus (see figure 1.1). The duplication of LTR leads to difficulties assembling the integrated provirus. Another hurdle is the insertion of retroviruses in repetitive or low complexity regions, which makes it almost impossible to retrieve the correct structure, length, and orientation of an ERV from short reads. Besides these limitations, we investigated historical samples which contained both highly degraded DNA and high amounts of contamination, resulting from exogenous DNA amplified during initial PCR cycles [142]. Even long-read sequencing would not cope with ancient DNA constraints as most fragments do not exceed an average length of 400 bp [143]. In this study, target enrichment efficiency was highest in SPEX with approximately 5 percent and lower in PEC and HC (0.55 percent and 0.01 percent, respectively; see table 2.1). In figure 2.4 it is shown that the vast number of fragments retrieved from the samples had a sequence length of around 50 bp. Since short-read sequencing is still less error-prone than long-read sequencing, utilization of short-read sequencing for analysis of ancient DNA samples is recommended.

One of the most resource-intensive processing steps during this analysis was the computation of sequence distances for integration sites, prior to clustering. We used NCBI BLAST+ for an all versus all sequence alignment. The basic local alignment search tool is a fast heuristics for alignments on nucleotide level. More efficient algorithms exist for protein alignments [144, 73, 74, 71]. Unfortunately, alignments on nucleotide level are either insufficiently sensitive or are based on optimal alignments, which are considerably slower and more memory-intensive [63, 64]. This field is still under development since sequencing is getting cheaper and massive amounts of data are produced which needs to be analyzed. One of the latest tools, NSimScan, which is more sensitive and faster than megaBLAST or BLAST+, was published after we finished this study [72]. Apart from improving the logic or algorithms of sequence alignments, significant progress was made within the last years by parallelization. Additionally, Graphics Processing Units (GPUs) became available as a general-purpose processing platform, which led to the development of adapted tools for GPU usage [70]. Despite increased memory consumption and a trade-off between computing and read/write, input/output operations, insufficient parallelization of the BLAST algorithm has been widely reported. Recently, it has been shown, that machine learning could close the gap for sequence similarity clusters on the nucleotide level [145].

In 2018, the koala genome was assembled, annotated and published [146]. It would have probably been considerably simpler to analyze the data using an assembled reference genome. One approach could have been to mask the koala retrovirus in the reference genome and map the gDNA from enriched integration sites to the masked genome.

We observed a bias towards the 5' integration site for HC which was previously described by Tsangaras et al. (2014) [123]. The other two target enrichment techniques were biased towards the 3' integration site. In later studies, we observed the phenomenon that both flanks of one integration site rarely had a similar number of amplification products. This aspect, however, remains to be explained.

In further studies, we examined that the target site duplications were much more variable than assumed in 2016. It has been shown that target site duplications can range from four to ten base pairs [2]. This may be a reason why we were not able to pair all integration sites. We based the pairing routine on the assumption that four nucleotides on each end of the gDNA fragment are duplicated, as only target site duplications of four base pairs were reported until that point [141]. Due to the length limitations resulting from DNA degradation, short-read sequencing, and LTR duplication we were not able to assemble corresponding proviruses for integration sites. This limits our ability to distinguish complete proviruses, recombined ERVs or degraded ERVs like solo-LTRs. Detecting around 1,200 integration sites from ten individuals would indicate that, on average, 60 ERVs were incorporated per individual. This number of KoRV integrations is consistent with recently published data [2, 146].

In conclusion, nowadays it would be recommended to combine PEC and HC for target enrichment of ERV, filter contaminants using an alignment to the viral reference and map the remaining reads to a masked reference genome. The approach described in this chapter also applies to other host species of which no reference genome is available.

Chapter 3

Sonication Inverse Polymerase Chain Reaction (SiP)

Retroviruses are unique as they require integration into the host genome to achieve viral replication. The process is accomplished through the use of specific viral enzymes that transcribe the retroviral RNA genome to DNA and integrate it into the host genome. The provirus subsequently exploits the host cells replication machinery, thereby obtaining all the necessary factors required for viral expression. The process is mutagenic, as the viral integration site represents a permanent alteration of the host DNA within an infected cell. Depending on the site of integration, adverse effects such as the disruption or deregulation of host genes (e.g. cell cycle and oncogenes) may be observed. Retroviruses have also been used as efficient vectors in gene therapy trials and across mutagenesis studies to identify genes involved in oncogenesis [147]. Early gene therapy trials, however, resulted in severe side effects associated with insertional mutagenesis by viral vectors [148]. The ability to characterize retroviral integration sites is therefore fundamental for understanding integration site preferences. Retroviral integration site profiles could serve as an assessment tool of host genome-pathogen interactions. Nevertheless, the comprehensive identification of retroviral integration sites is generally challenging, as similar sequences integrated into dozens or hundreds of genomic loci must be identified.

There are relatively few molecular methodologies for integration site identification that have been adapted to current genomic sequencing approaches. Presently, various polymerase chain reaction (PCR)-based strategies including linear amplification-mediated PCR (LAM-PCR) [149, 101], ligation mediated PCR (LM-PCR) [150], and splinkerette PCR [151] are some of the most-commonly used methods to retrieve viral integration sites. All three methods require the digestion of gDNA with a restriction endonuclease, the ligation of a linker cassette or adapter(s), and the amplification of the retroviral/host integration sites via primers that anneal to conserved regions of a retrovirus/viral vector and the ligated adapter(s). However, their reliance on restriction enzymes limits comprehensive retrieval of viral integration sites as they are dependent on the presence of specific recognition motifs that are unevenly distributed across the genome. A variant of LAM-PCR with increased sensitivity that circumvents the use of restriction enzymes (nrLAM-PCR) [102] recently showed increased efficiency as compared to its predecessor and is currently one of the most robust approaches in use for viral integration site retrieval [152]. However, given that it is based on standard PCR, it is limited to targeting one end of a proviral genome (5' or 3' end) at a time. Alternative target enrichment strategies such as Single Primer Extensions (SPEXs) [153], Primer Extension

Captures (PECs) [131] and Hybridization Captures (HCs) [132] have also been applied to study viral integration sites across modern [154] and historical samples [155, 123, 1]. Recent advances in molecular techniques have enabled the enrichment and sequencing of DNA fragments of up to 20 kb in length [156, 157, 158].

While showing promise, the development of long-fragment enrichment techniques is in its infancy and has not yet been used to explore viral integrations sites. Furthermore, the lengths of the obtainable sequences may in some cases be limited by the interaction between the biotinylated oligonucleotides and streptavidin coated magnetic beads used across these methods. This may result in reduced enrichment of longer oligonucleotides, likely due to steric hindrance [152, 156]. The use of restriction enzymes or biotinylated oligonucleotides in many of the previously mentioned PCR and enrichment-based methods reduces the recovery of longer genomic targets. In some cases, their inherent nature and the availability of commercially available kits (for hybridization capture) have made them ideally suited to the technological developments in short-read platforms of the last decade. However, with advances in long-read HTS there is a niche to adapt or develop methods that can simultaneously acquire viral integration sites in conjunction with viral integrant characterization at specific loci.

3.0.1 Laboratory Pipeline

Inverse PCR [159] is a variant of PCR that has historically been used to study retroviral integration sites [160, 126]. Its premise requires the fragmentation of gDNA followed by the intra-molecular circularization of DNA fragments. Inverted PCR primers designed end-to-end on conserved regions of a retroviral genome such as the long terminal repeat (LTR) are then used for targeted amplification of viral integration sites. Since its development, inverse polymerase chain reaction (iPCR) has fallen out of contemporary use. However, an adaptation of the method holds several advantages over current molecular approaches to characterize viral integration sites. In the present study, we describe SiP - a sonication-based iPCR strategy, coupled with Pacific Biosciences PacBio RSII platform. This technique was recently used to determine how recombination of the invading koala retrovirus (KoRV) and an older endogenous retroviral element occurs at the earliest stages of genomic invasion within the host species [2].

We evaluated SiP as a tool for comprehensive viral integration site retrieval in a high copy integration model using KoRV as an example. In this context, we (i) employed Covaris-based sonication to randomly fragment DNA and avoid the use of biased-cutting restriction enzymes, (ii) we tested and demonstrated adapter ligation deficiencies across DNA ligation experiments including those used in the generation of blunt end HTS DNA libraries, (iii) we used SiP to characterize proviral integration sites from sequences of up to 10 kb in length. The molecular technique and analytical pipeline proposed can be used to obtain any unknown sequence information flanking a known sequence and is, therefore, not limited to integration site analysis. SiP will, therefore, have broad applications in characterizing various mobile genetic elements, genome applications across taxonomy, as well as re-sequencing projects where deciphering structural variation (e.g. duplications, tandem repeats, and recombination) is challenging.

3.1 Methods

DNA Extraction, Fragmentation and End Repair [2]

The employed protocol has four steps:

1. DNA is extracted, randomly fragmented and end-repaired

2. the end-repaired DNA is then divided into two groups. One group had an adapter ligated before circularization (Adapter Ligation Group), while the other does not (Non-Adapter Ligation Group)

3. the two groups are circularized to produce closed circular DNA

4. long terminal repeat (LTR) and polymerase gene (*pol*) primed long fragment PCRs are performed on the closed circles

5. the resulting KoRV integration-enriched PCR products are PacBio sequenced and analyzed.

0 All quality control tests across the five steps above to determine DNA concentration and DNA size distribution were performed using the Qubit Fluorometer (High Sensitivity chemistry), as well as the 2200 TapeStation (Agilent Technologies) using Genomic DNA (gDNA) Screen Tapes.

Step 1, fragmentation: DNA extracts from "Bilyarra" (Pci-SN241) were quantified with the Qubit Fluorometer (high sensitivity chemistry) and the 2200 TapeStation (Agilent Technologies) using gDNA Screen Tape, showing DNA size to be primarily distributed around 50-60 kb. To produce DNA fragments of suitable size for this study (average of 3 to 4 kb), DNA extracts were diluted to 50 ng/μL in a final volume of 200 μL. The extracts were fragmented with a Covaris M220 in a miniTUBE Blue using the following settings (in parentheses): intensity (0.1-0.5), duty cycle (20%), cycles per burst (1000), total treatment time (600 seconds), temperature (20 °C). DNA fragmentation profiles of sheared DNA were further assessed on the TapeStation using a gDNA Screen Tape. After confirming that the size of the sheared DNA was between 2-7 kb, 42.5 μL of the sheared DNA was used in blunt-end reactions, run in triplicate using the Fast DNA End Repair Kit (Thermo Scientific). The products were purified using the QIAquick PCR Purification Kit (Qiagen, Hilden, Germany) and were again quantified using the TapeStation. The results across the triplicate runs were similar, thus they were pooled.

Step 2, Two complementary oligos were synthesized for the construction of the adapter (table 3.1). Each contained a 5' phosphate to facilitate subsequent blunt end ligation. The oligos were annealed together by following the Illumina sequencing adapter preparation procedure in a published protocol [161]. The Adapter Ligation Group was set up using a T4 DNA Ligase kit (5 U/μL) (EL0014, Thermo Scientific) with 5 μL of T4 DNA ligase buffer (10X), 5 μL of 50% PEG 4000 solution, 1 μL of Adapter (50 μM), 2.5 μL of T4 DNA Ligase, and 36.5 μL of blunt-ended DNA in a 50 μL total volume. Ligation was performed in a thermal cycler at 22 °C for 60 min, followed by enzyme inactivation at 65 °C for 10 min. Ligation products were purified using the Agencourt AMPure XP PCR Purification system (Beckman Coulter GmbH, Krefeld, Germany). A bioinformatic assessment of DNA size among the purified products from the Adapter Ligation Group

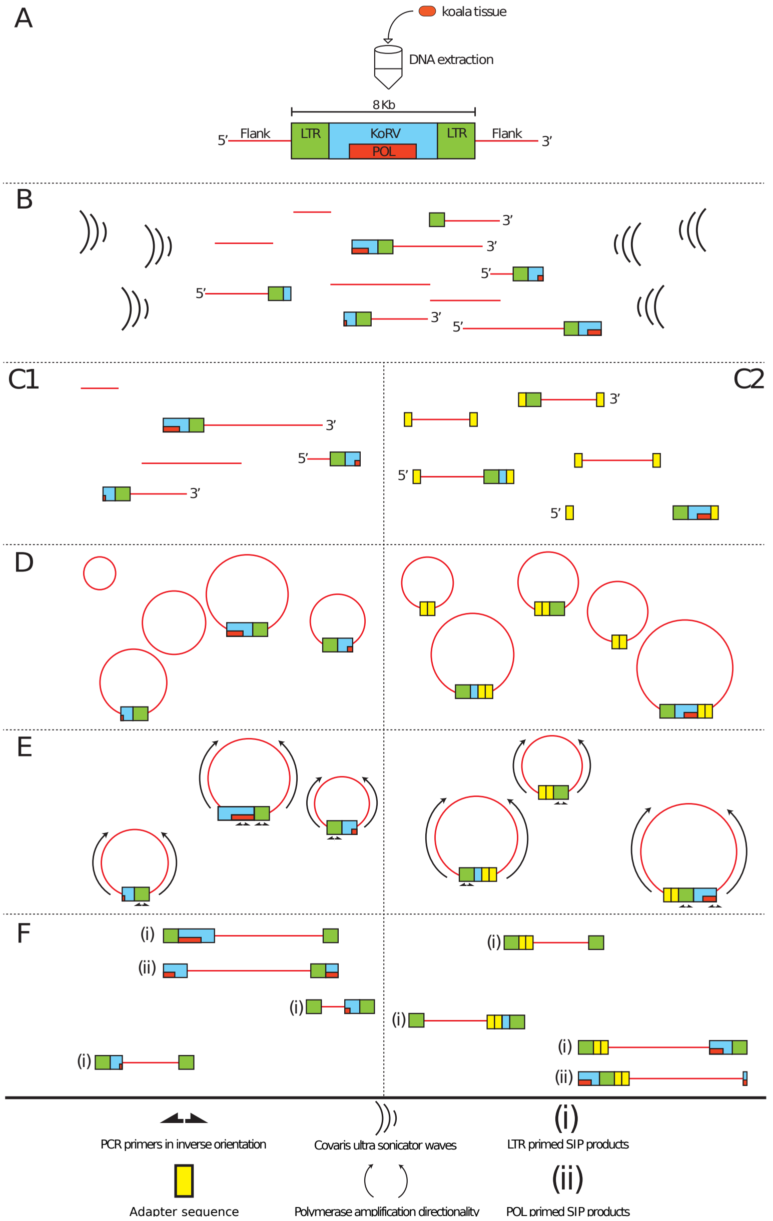


Figure 3.1: Sonication Inverse PCR Scheme [3]

(A) The KoRV provirus, which is integrated into the koala genomic DNA is illustrated with typical LTR regions (green box) flanking the retroviral genes (blue box). Note: Only the approximate location of the *pol* gene (red box) is indicated diagrammatically for simplicity. (B) Koala genomic DNA was fragmented to an average length of 2-7 kb using ultra-sonication. The fragmented DNA was then blunt-end repaired and phosphorylated. (C) The sample was subsequently divided in two aliquots, a non-adapter group (C1) and an adapter Group (C2). The non-adapter group was not modified in any way prior to circularization, whereas the adapter group had an identical adapter sequence (yellow box) ligated on either end of the DNA molecule for assisted interpretation of the inverted amplicon sequences following circularization and amplification. (D) Both the adapter and non-adapter groups were circularized resulting in circular DNA templates. (E) Circularized DNA templates were amplified with two primer sets that target the *pol* and LTR regions of KoRV. Circularized templates without these primer-binding sites do not amplify. (F) Amplified products were inverted with the primer-binding site located on the flanks of the amplicon. Two types of PCR product were generated: (i) PCR products amplified by the LTR primers and (ii) PCR products amplified by the *pol* primer.

showed a similar size distribution to the blunt-ended DNA from the Non-Adapter Ligation Group (figure 3.4).

Step 3, Circularization: to identify the optimal ligation conditions for subsequent iPCR, ligations were performed using a series of varying (total) input blunt-ended DNA. The following amounts were tested : 5 ng, 10 ng, 15 ng, 25 ng, 30 ng, 40 ng, 50 ng, 75 ng, and 100 ng , each in a 50 μL ligation reaction. Ligation reactions used a T4 DNA Ligase kit (5 U/μL) (Thermo Scientific) with 5 μL of T4 DNA Ligase Buffer (10X), 5 μL of 50% PEG 4000 solution, 2.5 μL of T4 DNA Ligase and an amount of blunt-ended DNA (as indicated by the series), and molecular biology-grade water to a total volume of 50 μL. Ligation was performed in a thermal cycler at 16 °C for 16 hours followed by enzyme inactivation at 70 °C for 5 min. A non-template circularization control (control 1) was run simultaneously for each gradient. Given the small starting amounts of DNA, all ligations for both groups and control 1 were performed in triplicate to minimize bias. Triplicate results of the same input amount showed high similarity with each other based on DNA size assessments and were subsequently pooled. The ligation products were measured on the TapeStation, showing a 2 kb size shift towards higher molecular weight compared to the blunt-ended DNA, a sign of transformation of DNA structure from linear to circular. Results for each of the ligation reactions using the same amount of DNA were similar in profile as measured on the TapeStation, so the products of each three were pooled together. A partial ligation (circularization) gradient (25 ng, 30 ng, 40 ng, 50 ng, 75 ng) was rerun to test reproducibility, showing comparable results.

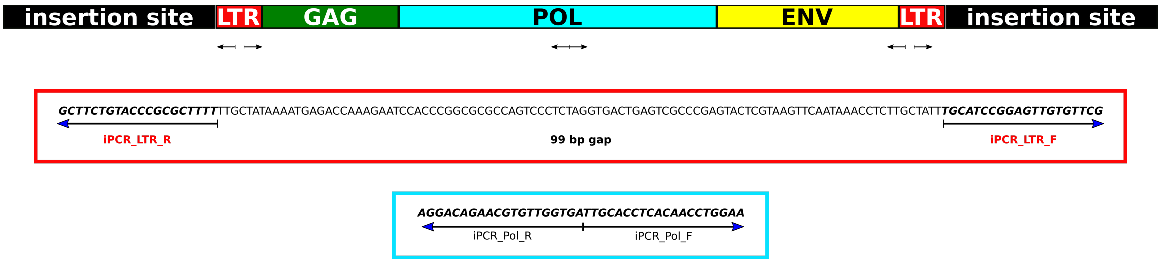

insertion site
LTR
GAG
POL
ENV
LTR
insertion site
GCTTCTGTACCCGCGCTTTTTTGCTATAAAATGAGACCAAAGAATCCACCCGGCGCGCCAGTCCCTCTAGGTGACTGAGTCGCCCGAGTACTCGTAAGTTCAATAAACCTCTTGCTATTTGCATCCGGAGTTGTGTTCG
iPCR_LTR_R
99 bp gap
iPCR_LTR_F
AGGACAGAACGTGTTGGTGATTGCACCTCACAACCTGGAA
iPCR_Pol_R
iPCR_Pol_F

Figure 3.2: Inverse PCR Primer

Two sets of primers were designed targeting a conserved region of the KoRV LTR (red rectangle) and a conserved region of POL (blue rectangle). Arrows indicate the primer binding sites associated to the complete viral structure and relative to the domains. Primer binding sites are shown relative to the scheme of figure 1.1. While the LTR primers are designed with a gap of 99 bp between them, the POL-primers were designed side by side.

Step 4, Long iPCR: KoRV proviral genomes were downloaded from GenBank (accessions: KF786280, KF786281, KF786282, KF786283, KF786284, KF786285, KF786286, AB721500, KC779547) and were aligned using the MAFFT plugin in Geneious version 7.1.7 using default settings [162]. For the iPCR, two sets of primers were designed using Primer3Plus software [163] targeting a conserved region in the middle of the KoRV LTR, and a conserved region in the middle of POL shown in figure 3.2 and table 3.1.

To avoid loss of circularized DNA during purification, circularization products were

Table 3.1: List of Primers and Oligonucleotides [3]

Primer/Oligonucleotide	Sequence (5'-3')
iPCR_LTR_F	TGCATCCGGAGTTGTGTTCG
iPCR_LTR_R	AAAAGCGCGGGTACAGAAGC
iPCR_POL_F	TTGCACCTCACAACCTGGAA
iPCR_POL_R	TCACCAACACGTTCTGTCCT
iPCR_adaptor_F	Pho-CTGAGTCGGAGACACGCAGGGATGAGATGG
iPCR_adaptor_R	Pho-CCATCTCATCCCTGCGTGTCTCCGACTCAG

Pho indicates a phosphate group

used directly as templates for the iPCR without purification [164]. A total of 10 ng (as quantified by the TapeStation) from each product in the circularization gradient was used as template in a separate iPCR. The template was amplified using the MyFi Mix (Bioline GmbH, Luckenwalde, Germany) with thermal cycling conditions of an initial denaturation step at 95 °C for 1 min 30 sec; followed by 35 cycles at 95 °C for 20 sec, 59 °C for 20 sec and 72 °C for 5 min; final extension of 2 min at 72 °C. The same PCR conditions were used across both LTR and *pol* amplifications using MyFi Mix (Bioline GmbH, Luckenwalde, Germany). Additionally, two controls for every PCR amplified gradient were run including; a non-template PCR control (control 2) and a linear control of fragmented-blunt-ended genomic koala DNA (control 3).

The products were purified using the QIAquick PCR purification kit (Qiagen, Hilden, Germany), and concentration and DNA profiles were measured on the TapeStation. The optimal circularization product of each gradient was chosen by considering (i) the DNA amount per micro-liter of iPCR product in the 2-7 kb range, (ii) the average length distribution between 600 bp-7 kb range, and (iii) the percentage of DNA was within the 2-7 kb range. Based on these criteria, 40 ng of input DNA (conc. 0.8 ng/μL in circularization) for both the Adapter Ligation Group and the Non-Adapter Ligation Group were chosen as the optimal circularization product. To test whether increasing the template amount (circularization product) for iPCR would affect the length distribution of the PCR product, a series of template amounts (2 ng, 6 ng, 10 ng, 14 ng, 16.8 ng, which is the amount of DNA in a maximum input volume of 21 μL) from the two optimal circularization products were used for iPCR using same kit and protocols described above. All products were purified using the QIAquick PCR purification kit (Qiagen, Hilden, Germany) and were measured on the TapeStation. The optimal iPCR product amount in the series was chosen based on the three criteria above and on the overall distribution of the fragment size peaks determined by the TapeStation measurement. Following this analysis, the iPCR with optimal template amount (6 ng) was repeated twice. These three optimal products were then pooled for PacBio sequencing to minimize clonal PCR bias.

Step 5, PacBio library preparation, sequencing, and data curation: two pools of PCR products, consisting of either the Adapter Ligation Group or the Non-Adapter Ligation Group, were submitted to the Max Delbrück Center for PacBio library construction and sequencing. Both PCR product pools were purified using AMPure XP beads (Beckman Coulter), first at a concentration of 0.4X followed by a subsequent purification of the supernatant at 0.6X. The resulting four samples from these purifications were then prepared as sequencing libraries using the PacBio (Pacific Biosciences, Menlo Park, CA) 5 kb template preparation protocol and the SMRTbell^{TM} Template Prep Kit 1.0 following the manufacturer's guidelines. The libraries were estimated at an average length of

1,600 bp and 3,500 bp for the short and large insert libraries respectively using the 2100 Agilent Bioanalyzer and the 1200 DNA chemistry (Agilent Technologies). Sequencing on the PacBio RSII platform was done, using the MagBead Standard protocol, C4 chemistry and P6 polymerase on a single v3 Single-Molecule Real-Time (SMRT) cell with a 1×180 min movie for each library (a total of four libraries – Adapter Ligation Group: short and long insert libraries and Non-Adapter Ligation Group: short and long insert libraries). Reads from the insert sequence were processed within the SMRTPortal browser (minimum full pass = 1; and a Minimum Predicted Accuracy of 90; read length statistics shown in table 3.2).

Table 3.2: Sequencing Statistics SiP PacBio

Description	IS [bp]	Adapter	# reads	Quality	∅ passes	∅ length
Adapter long	5000	NO	28,983	0.9847	10	2,347
Adapter short	2000	NO	31,794	0.9855	19	1,061
Non-Adapter long	5000	YES	24,076	0.9852	11	2,175
Non-Adapter short	2000	YES	26,910	0.9884	18	1,255

Blunt-End 454 DNA Library Construction and PacBio Sequencing Using Illumina Adapters

The same fragmented koala DNA extract decribed above was used to build a blunt-end HTS DNA library. Library construction was based on the general principal developed for 454 sequencing [165], which was later modified using Illumina adapters and is one of the standard protocols for generating Illumina libraries [161].

The following modifications were made to the previously described Illumina adaptation: (i) All SPRI bead purification steps were substituted with spin column modifications (QIAquick PCR purification kits, Qiagen), (ii) A final adapter concentration of 1 μM was used to build the libraries - the same concentration as the Adapter Ligation Group (iii), the fill-in reaction procedure was performed at 65 °C for 20 min, (iv) all columns were incubated at 37 °C for 5 min prior to elution, and (v) the final purification following the fill-in reaction was omitted. Despite adaptation and sequencing on an Illumina platform, to retain a consistent nomenclature, we refer to this DNA library as the blunt-end 454 DNA library as previously described elsewhere [166]. Successive SiP steps followed the same procedures outlined above including: blunt-ending, inter-molecular circularization, and amplification using the same LTR primers and conditions. A circularization concentration of 0.8 ng/μL (40 ng total DNA input) was used as it was previously determined to be the optimal ligation concentration for the sample. The 454 DNA library was subsequently built into a PacBio sequencing library as described above and purified using an AMPure XP bead (Beckman Coulter) concentration of 0.9X. Subsequently, the blunt-end 454/PacBio DNA library was sequenced on half a SMRT cell using the MagBead standard protocol with a data collection time of a 1×240 min movie on the PacBio RSII platform.

3.1.1 Bioinformatic Analysis

Calculating KoRV Sequence Enrichment

A KoRV reference database was created by downloading genomes of KoRV-A (KF786280) and KoRV-B (KC779547) from GenBank. All four datasets were searched for KoRV. All reads of insert (ROIs) were aligned to the KoRV references using megaBLAST [67, 68] with default settings. A subsequent analysis using an e-value of 10^{-5} showed no change to the number of aligned sequences. KoRV positive ROI were aligned to the NCBI nt database (NCBI-GenBank Flat File Release 220.0) using megaBLAST with an e-value restriction of 10^{-5}. Results were visualized using KronaTools [167]. The same alignment and visualization process was applied to sequences, determined as off-target reads, which could not be aligned to KoRV.

Adapter Search

All ROIs were separately aligned to KoRV domains (LTR, *gag*, *pol*, *env*), primer sequences and adapter sequences (BLASTn). The results were parsed to identify the most common structural variants of SiP reads. Adapter sequences were validated by a minimum alignment length of 25/30 bp, 25/33 bp, 25/34 bp, depending on the length of the oligonucleotides used to construct each adapter. Primer sequences were validated by a minimum alignment length of 15/20 bp. Eight major groups of structural variants of SiP reads were constructed and evaluated by counting the occurrence of distinct motives described in figure 3.10.

3.2 Results

3.2.1 Development and Testing of SiP

A visual summary of SiP is presented in figure 3.1. The method requires the initial circularization of fragmented-blunt-ended DNA, followed by targeted amplification of KoRV using primers to the long terminal repeat (LTR) and the polymerase gene (*pol*). SiP provides an alternative to previously described methods in a simplified workflow (figure 3.3), enabling the dual characterization of integration sites and proviral sequences through long read HTS.

The development of an optimized workflow for SiP required the testing of several intrinsic factors such as; (i) establishing the optimal iPCR conditions and (ii) the implementation of various controls throughout SiP's workflow. Pooled triplicates of circularized koala gDNA were used as a template for SiP. TapeStation readings of iPCR products indicated the presence of large peaks beyond the size of the initial fragmented gDNA. Optimization of the SiP cycling conditions including a reduction of the polymerase extension times and the number of PCR cycles (data not shown) reduced the formation of these artifacts. We suggest that this was due to over-amplification of a low amount of template DNA, which resulted in the formation of large DNA concatemers.

Three controls were used to monitor SiP performance (figure 3.3). Control 1 consisted of a non-template circularization blank to monitor the introduction of DNA contamination at the circularization step. The assessment consisted of taking Control 1 through the whole experimental workflow. Control 2 consisted of a non-template control of the iPCR reaction for each gradient and group. TapeStation assessments of purified products

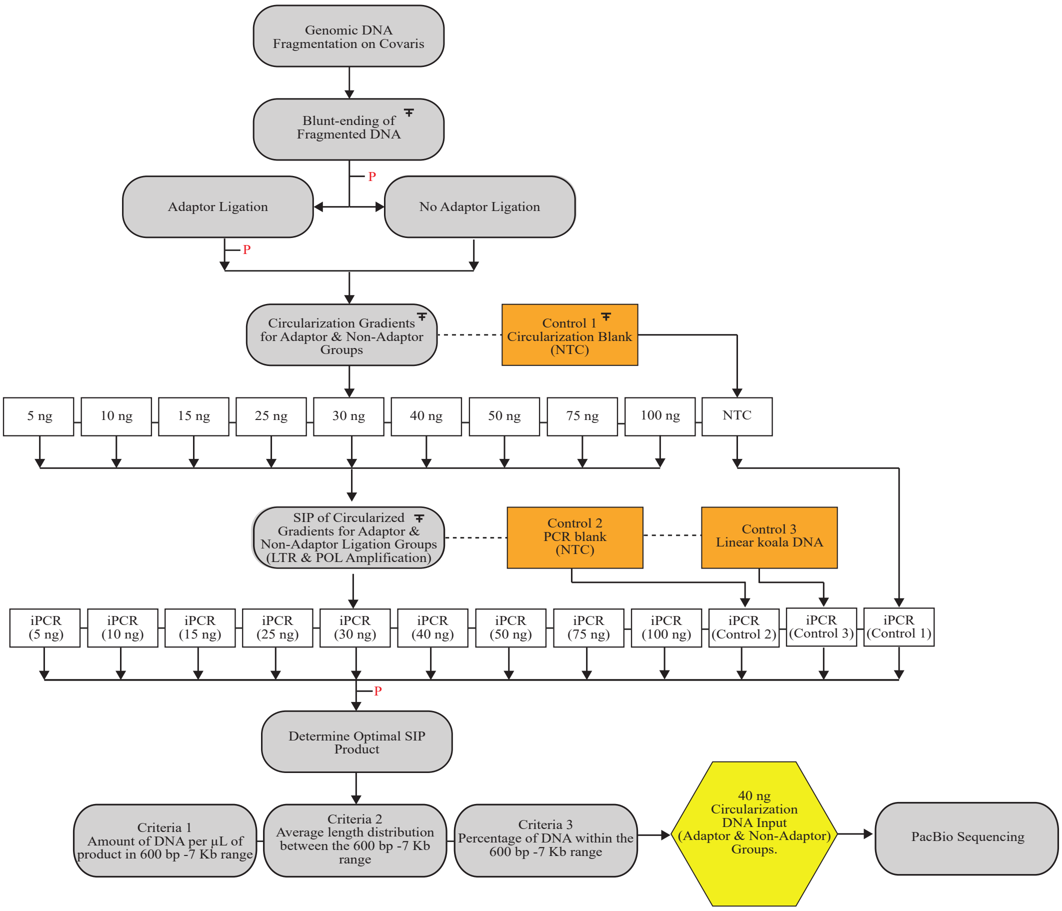


Figure 3.3: Experimental Workflow of Sonication Inverse PCR [3]

Abbreviations used in the figure include: SiP: Sonication Inverse PCR, P: purification, ∓: Triplicate reactions, NTC :non-template control, iPCR: inverse Polymerase Chain Reaction. Grey rounded rectangular boxes denote important steps in the workflow, white rectangles indicate gradient steps and orange rectangles are controls. Workflow: Purified genomic koala DNA was fragmented to an average length of 3-4 kb. The extract was then blunt-ended and divided into either an Adapter Group, where an adapter was ligated on either end of the DNA fragment pool, or a Non-Adapter Group. A circularization gradient of total DNA was then used to test self-ligation efficiency for both groups. Inverse PCR was performed on all gradient points for both groups using two different sets of primers (LTR and *pol*) and the purified amplicons were measured on a TapeStation. Three criteria were used from the TapeStation profiles to assess the optimal amplification gradient from each group for PacBio sequencing (yellow hexagon). Three controls were used throughout the experiment. Control 1: A non-template water control was run all the way through the experimental workflow starting from the circularization procedure. This control was used to monitor for DNA contamination from the circularization step. Control 2: A second non-template water control was run during the iPCR step and was used to monitor DNA contamination introduced during PCR setup. Control 3: A linear DNA control was used to assess PCR amplification of non-circularized (linear) gDNA template.

from Controls 1 and 2 produced no visible amplification products, thereby confirming the absence of DNA contamination. Control 3 consisted of linear control of fragmented blunt-ended koala gDNA. TapeStation readings displayed some minor observable amplification peaks, suggesting that un-circularized (linear) DNA can be amplified with primers in inverse orientation. Standard PCR amplification could occur if more than one provirus is located in close proximity to another in the host genome, where the forward LTR primer at the 3' end of a provirus primes with the reverse LTR primer of a 5' provirus or vice versa. Non-circularized DNA may also be primed by a single PCR primer to produce amplicon products through a linear (non-exponential) amplification.

3.2.2 Evaluating SiP's Library Length Distribution, KoRV Sequence Enrichment and Off Target Enrichment

Central to SiP's application is the inter-molecular circularization of the 5' and 3' ends of a DNA molecule. An important consideration of this process is that upon circularization the ends of the DNA molecule will be obscured and may complicate analysis. To circumvent this issue, we tested the effect of adding an adapter by dividing the experiment into two groups (an Adapter Group and a Non-Adapter Group) to compare the eventual performance between the two (figures 3.3 and 3.1). The premise behind the Adapter Group was to mark the sheared boundaries of the blunt-ended DNA fragments, important for the biological interpretation of iPCR products.

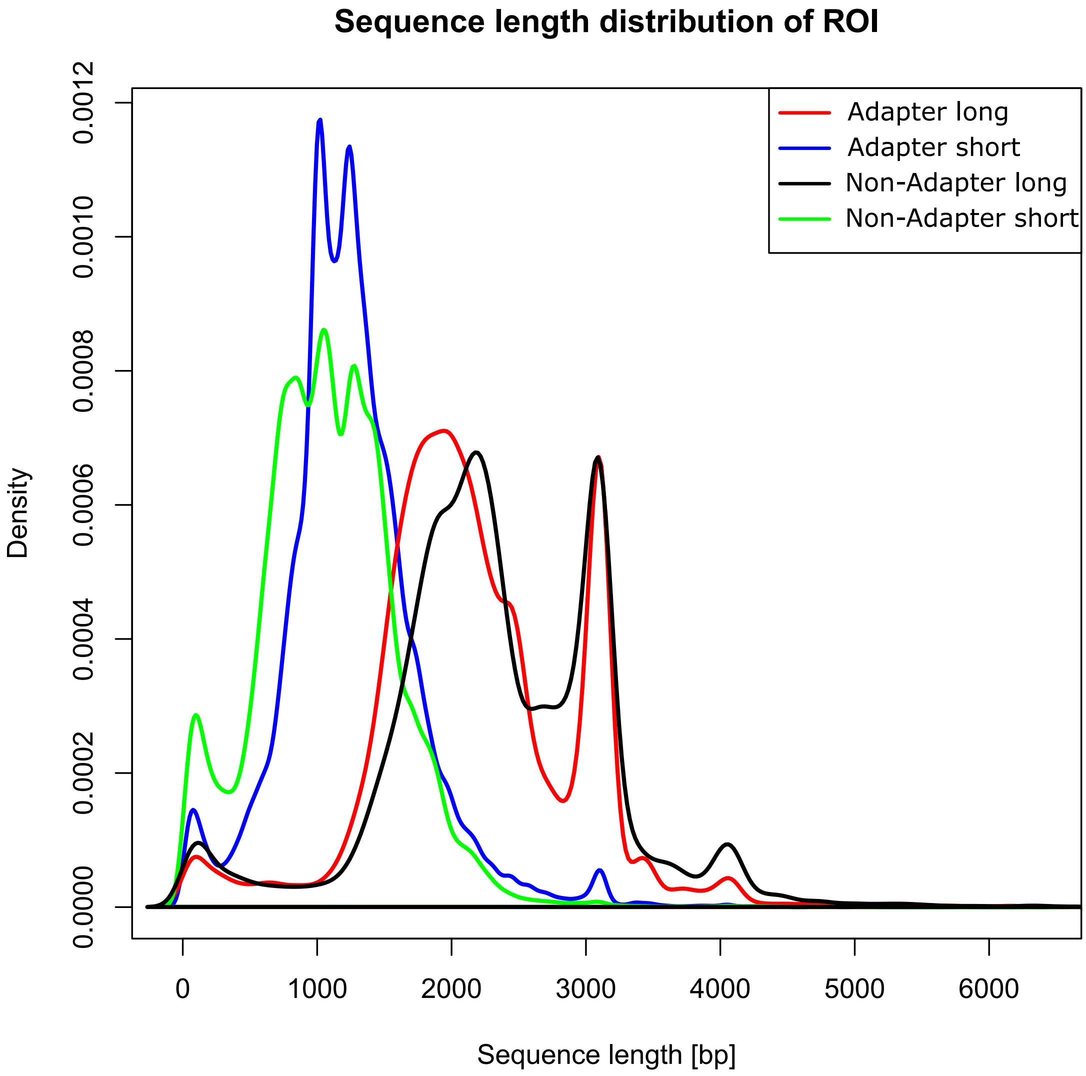


Figure 3.4: Reads of Insert (ROI) Sequence Length Distribution [3]

Sequence length distributions of reads from adapter long group (red), adapter short group (blue), non-adapter long group (black) and non-adapter short group (green). As expected, the long fragment groups (red and black) are showing a shift in length distribution to longer sequences with two peaks at approximately 2 kb and 3 kb, while most sequences from the short fragment groups (blue and green) have a length of approximately 1 kb.

As an adapted iPCR technique, it was initially unclear whether inter-molecular circularization of DNA fragments is length limited. LTR and *pol* amplicons from each of the Adapter and Non-Adapter Groups were first pooled and built into two PacBio libraries. Each library was then size-selected (referred to as long and short insert libraries – refer to figure 3.4) using two different length cut-offs, (refer to methods for details) and were compared to test the upper and lower length limits of the amplified products. As a measure of enrichment, all four PacBio sequence datasets were evaluated for KoRV-like sequences using BLAST at the nucleotide level. The analysis showed an exceptionally high enrichment of KoRV-like elements. Especially notable were the non-adapter long and non-adapter short datasets, which yielded total KoRV enrichment rates of 94 percent and 95 percent of all sequenced reads respectively (table 3.2.2). In contrast, the adapter (long and short) datasets had a lower total enrichment rate of 82 percent and 63 percent respectively. The highest KoRV enrichment for sequences longer than 1,000 bp derived from the two long insert libraries at 96 percent (non-adapter long) and 97 percent (adapter long). While the shorter datasets showed reduced enrichment of 58 percent (non-adapter short) and 77 percent (adapter short), the enrichment of KoRV sequences across the four datasets exhibited a mean alignment length of 1,111-2,396 bp. As expected, the longest KoRV homologous sequences were identified in the adapter long (9,864 bp), and the non-adapter long (9,590 bp) insert libraries. Our results indicate that the inter-circularization process can readily produce sequenceable amplicons of interest of approximately 60 bp to 10 kb in length.

Table 3.3: KoRV Sequence Enrichment Using SiP and PacBio RSII Sequencing

File	**Read count**	**Total nucleotides**	**<100 bp**	**>1000 bp**
Adapter long	24,076	52,396,194	249	22,962
Adapter short	26,910	33,811,389	395	19,975
Non-adapter long	28,983	68,056,034	390	27,435
Non-adapter short	31,794	34,423,855	925	18,021
File	**Min length**	**Max length**	**Mean**	**Median**
Adapter long	16	9,865	2,176	2,090
Adapter short	16	6,293	1,256	1,227
Non-adapter long	16	9,591	2,348	2,266
Non-adapter short	15	6,632	1,083	1,070

A breakdown of the off-target (non-KoRV) sequences (figures 3.5,3.6,3.7 and 3.8) using Krona [167] analyzed at the nucleotide level showed that between 51 and 68 percent of the non-KoRV sequences from the four datasets showed high similarity to the Tammar wallaby (*Notamacropus eugenii*) [168], the species most-closely related to the koala of which an assembled reference genome is available. The second largest fractions (16-28%) matched the koala genome (*Phascolarctos cinereus*) [146]. Taken together, between 79 to 84 percent of off-target reads were similar to wallaby or koala sequences. In addition, approximately 6-10 percent of the sequences were assigned to other eukaryotes, notably extant marsupials such as the Tasmanian devil, platypus, and opossum, while only a fraction of reads (0-0.9%) could not be assigned across the four datasets. Despite a search against the entire nucleotide database, the analysis of the reads yielded no identifiable bacterial sequences. A re-analysis of the off-target sequences at the protein level produced a comparable result to that of the nucleotide analysis (data not shown).

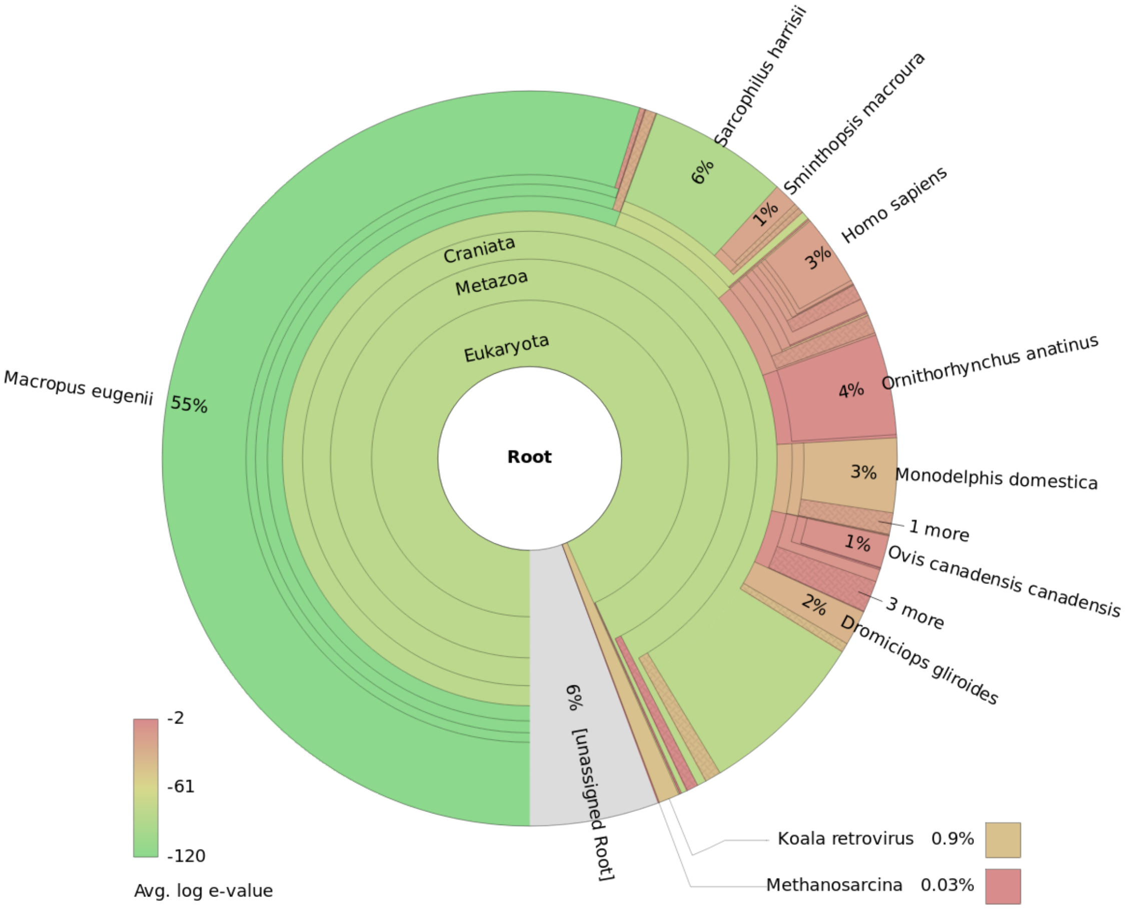


Figure 3.5: Adapter Long Off-Targets

The majority (55%) of adapter long off-target reads show a high similarity to the tammar wallaby genome. Six percent of the reads were similar to the genome of the tasmanian devil. Contamination could be expected from reads homologous to *Homo sapiens* (3%) and *Ovis canadensis*. An amount of 6 percent of the sequences could not be assigned to any reference in the Genbank database; 0.9 percent of the sequences were still to the koala retrovirus.

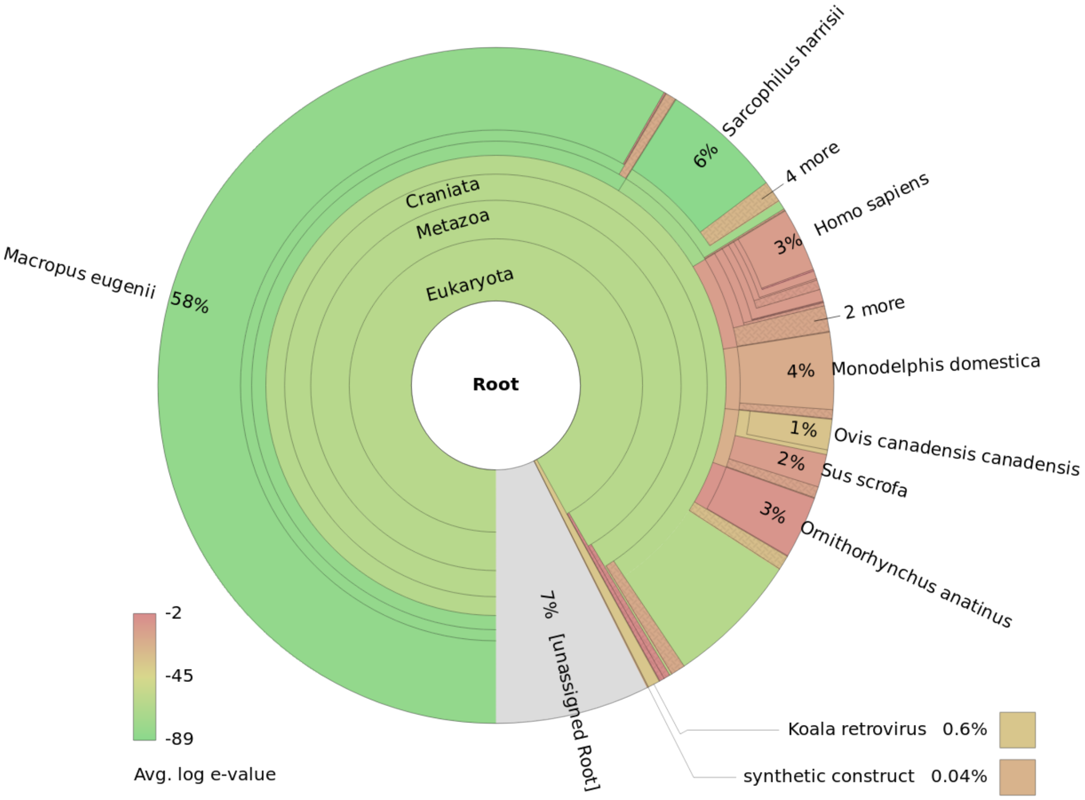


Figure 3.6: Adapter Short Off Targets

The vast amount (58%) of adapter short off-target reads show a high similarity to the tammar wallaby genome. 6 percent of the reads were similar to the genome of the tasmanian devil. Contamination could be expected from reads homologous to *Homo sapiens* (3%), *Sus scrofa* (2%) and *Ovis canadensis* (1%). An amount of 7 percent of the sequences could not be assigned to any reference in the Genbank database. 0.6 percent of the sequences were assigned to the koala retrovirus.

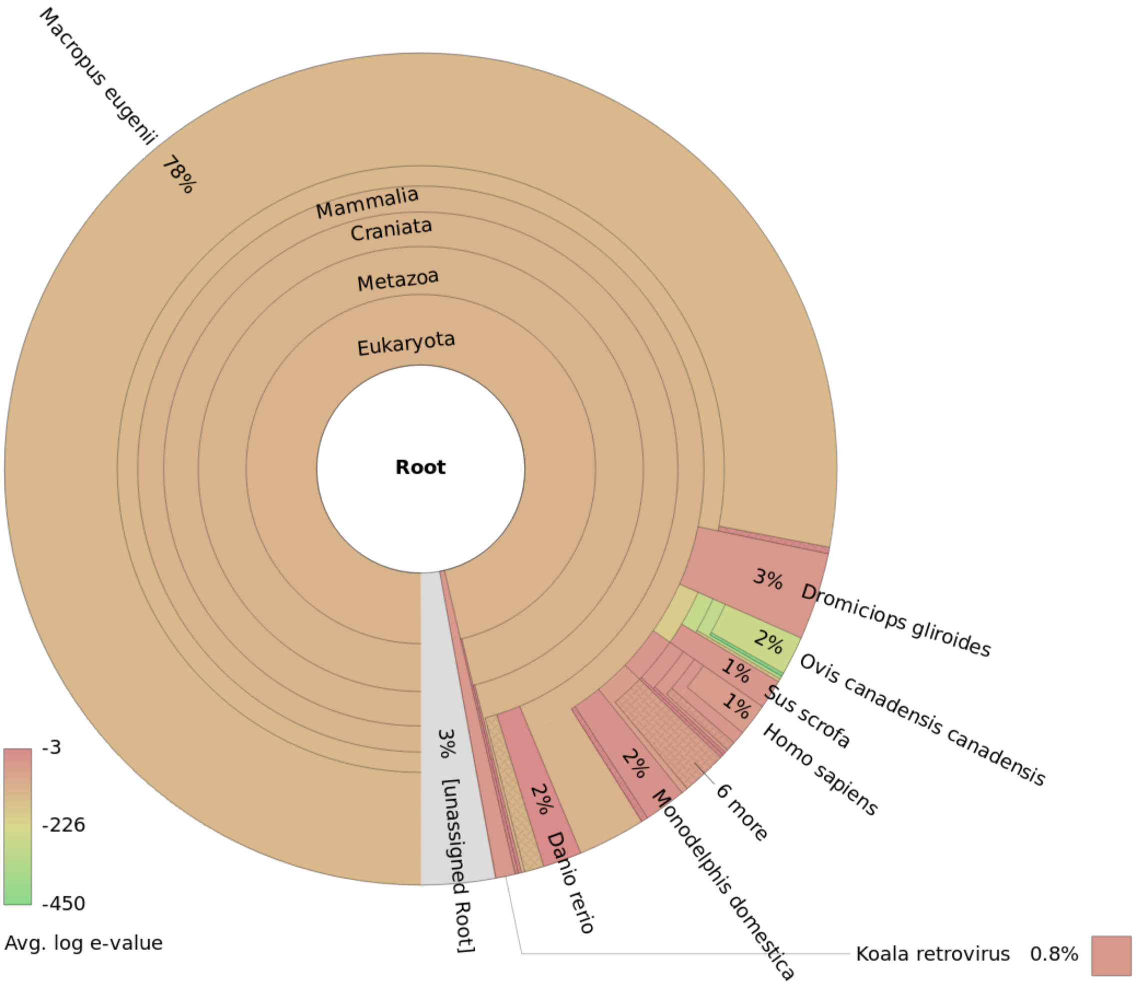


Figure 3.7: Non-Adapter Long Off Targets

The significance (determined by log e-value) of homologous references to the non-adapter long off-target reads is the lowest for all compared datasets. 78 percent of the sequences show highest similarities to the tammar wallaby genome. 6 percent of the reads were similar to the genome of the tasmanian devil. Contamination could be expected from reads homologous to *Ovis canadensis* (2%), *Homo sapiens* (1%) and *Sus scrofa* (1%). An amount of 3 percent of the sequences could not be assigned to any reference in the Genbank database. 0.8 percent of the sequences were still assigned to the koala retrovirus.

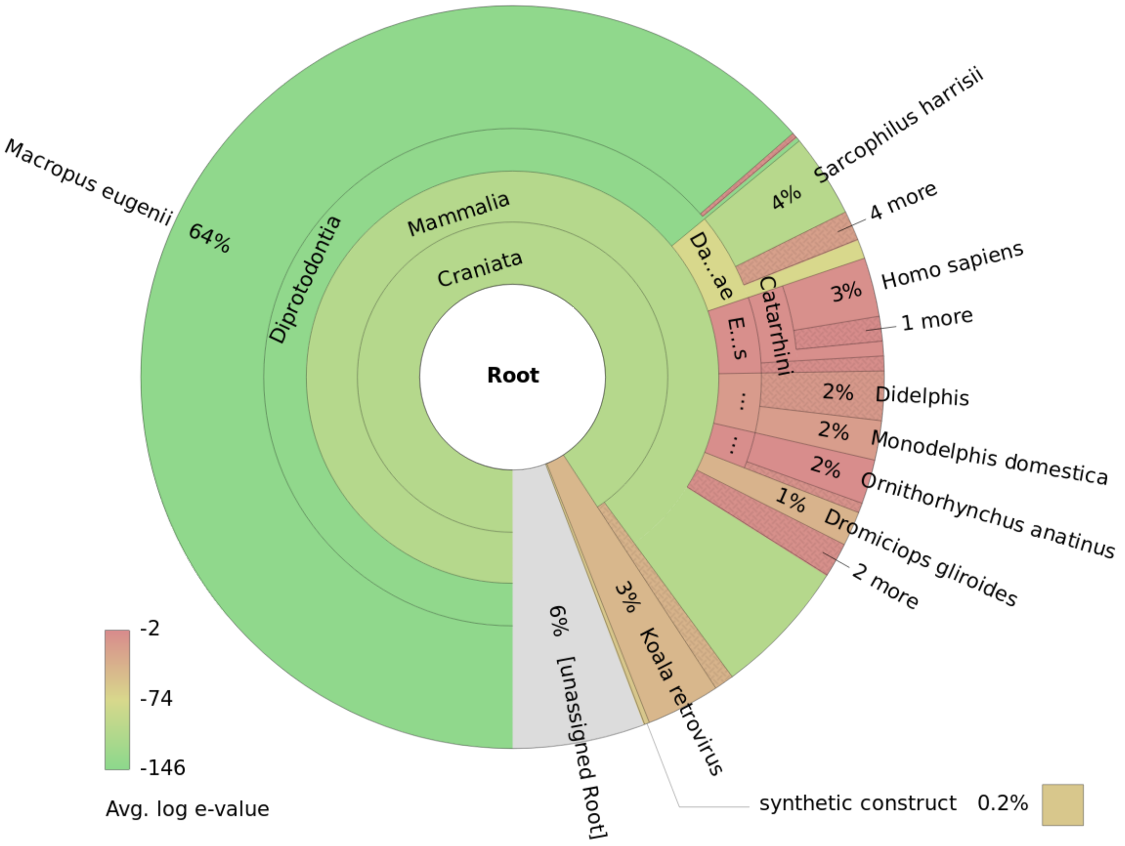


Figure 3.8: Non-Adapter Short Off Targets

64 percent of the adapter non-short off-target reads show a high similarity to the tammar wallaby genome. 4 percent of the reads were similar to the genome of the tasmanian devil with high significance. Contamination could be expected from reads homologous to *Homo sapiens* (3%). An amount of 6 percent of the sequences could not be assigned to any reference in the Genbank database. In the non-adapter short off-target dataset with 3 percent of the sequences the highest amount from all datasets were still assigned to the koala retrovirus.

3.2.3 SiP Blunt-End Adapter Ligation Efficiency and Blunt-End 454 DNA Library Adapter Ligation Experiment

We first assessed the ligation efficiency of the adapter in the four standard SiP datasets (adapter and non-adapter - long and short libraries). As expected, no adapter sequences were identified in either of the two datasets without adapters. However, adapter ligation enrichment for the two datasets with adapters was low, with the highest percentage of filtered reads with two adapters occurring in the adapter long dataset (4%), whereas the adapter short dataset had an adapter ligation efficiency of 2 percent. Therefore, the majority of reads (approximately 86-90 percent) in both standard adapter SiP datasets did not contain any adapter sequences (table 3.4). The data from the two standard SiP

Table 3.4: Adapter Counts Across Four SiP Datasets and Blunt-End 454 DNA Library (with Illumina Adapters) Dataset

		Filtered Reads with Two Adapters		Filtered Reads with > Two Adapters		Filtered Reads with No Adapters		Filtered Reads with Other Configurations^D	
Dataset	Total Filtered Reads^A	Total^B	Perc.^{C}	Total^B	Perc.^{C}	Total^B	Perc.^{C}	Total^B	Perc.^{C}
Adapter long	24076	1026	4	912	4	20791	86	1347	6
Adapter short	26910	607	2	722	3	24270	90	1311	5
Non-adapter long	28983	0	0	0	0	27059	93	1924	7
Non-adapter short	31794	0	0	0	0	28328	89	3466	11
Blunt-end 454 DNA library	18327	4436	24	562	3	10745	59	2584	14

A total number of reads that passed filtering criteria;
B total number of filtered reads with adapters or no adapter sequences;
C percentage of reads with adapters or no adapter sequences;
D total filtered reads with other adapter configurations (e.g. reads with two identical adapters) did not pass alignment criteria

datasets with the incorporation of an adapter (long and short insert libraries) suggests that blunt-end adapter ligation is an inefficient process and prompted us to test the efficiency of blunt-end Illumina adapter ligation (P5 and P7) when creating a 454 DNA library. The blunt-end 454 DNA library was generated from sheared koala gDNA using a variation of the Meyer & Kircher protocol [161]. The blunt-end 454 DNA library was then subjected to the same circularization and KoRV LTR iPCR priming procedures as previously described.

Importantly, the experimental approach amplifies circularized KoRV-454-DNA-library template regardless of whether blunt-end adapters are ligated to the ends of the DNA molecule (figure 3.9). By priming the PCR in the KoRV LTRs, a similar count of the enriched KoRV DNA molecules with and without Illumina P5 and P7 adapters attached could be calculated. An analysis of the sequence data from the 454 DNA library dataset indicated that only 24 percent of the reads had two adapter sequences in the correct orientation (table 3.4). In contrast, 59 percent of reads had no adapters, 3 percent of reads had more than two adapters and 14 percent had the same adapter attached. Overall, approximately 76 percent of the DNA within a blunt-end 454 DNA library is therefore not sequenceable as it is lacking the primer binding site for sequencing due to the incorrect number of ligated adapters (one or more than two), or the non-directional blunt-end ligation of the same adapter to a DNA library molecule.

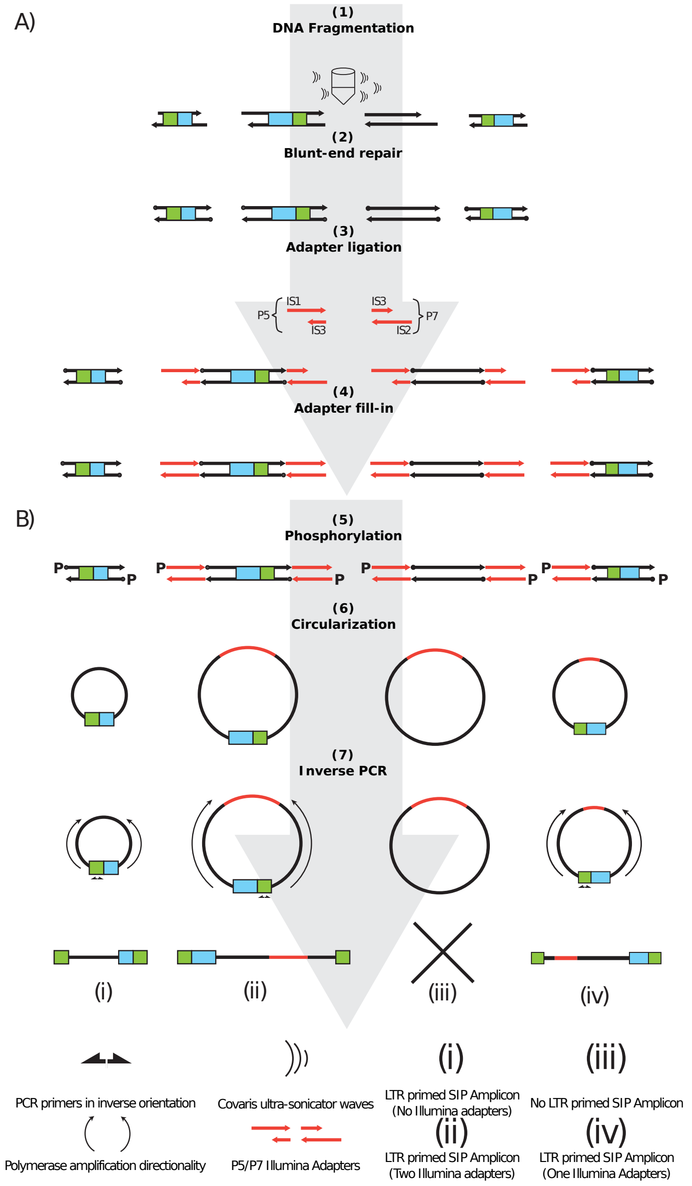

A)
(1)
DNA Fragmentation
(2)
Blunt-end repair
(3)
Adapter ligation
P5
IS1
IS3
IS3
IS2
P7
(4)
Adapter fill-in
B)
(5)
Phosphorylation
P
P
P
P
P
P
P
P
(6)
Circularization
(7)
Inverse PCR
(i)
(ii)
(iii)
(iv)
PCR primers in inverse orientation
Polymerase amplification directionality
Covaris ultra-sonicator waves
P5/P7 Illumina Adapters
(i)
LTR primed SIP Amplicon (No Illumina adapters)
(ii)
LTR primed SIP Amplicon (Two Illumina adapters)
(iii)
No LTR primed SIP Amplicon
(iv)
LTR primed SIP Amplicon (One Illumina Adapters)

Figure 3.9: Blunt-End 454 DNA Library Adapter Ligation Experimental Workflow [3]

A) Blunt-end 454/PacBio DNA library construction: (1) Genomic koala DNA, indicated by black lines with arrows, was sonicated to an average length of 3-4 kb. KoRV provirus (green box represents the LTR region, while the blue box indicates internal KoRV genes *gag*, *env* and *pol*), which is integrated into the koala genome, was also fragmented. (2) The sheared genomic DNA was repaired and blunt-ended. (3) Two different adapters were ligated to the ends of the blunted molecules. (4) A fill-in reaction repaired nicks and filled in lagging adapter strands.

B) SiP Procedure: (5) DNA library was phosphorylated (P), which added 5' phosphate and 3' hydroxyl groups. (6) Inter-molecular circularization of the DNA library ensues. Note: Circles represent double stranded DNA. The reaction will circularize all DNA library molecules, regardless of number of adapters ligated to distal ends (adapters are denoted by a red line within circle). (7) Circularized library was amplified using inverted KoRV LTR primers. Only circularized template with a KoRV LTR could amplify. Amplicons were built into PacBio libraries and sequenced on the PacBio RSII platform.

3.2.4 SiP Structure Variations

Bioinformatic analysis of the sequence reads from the five datasets (adaptor and non-adaptor - long and short datasets and the 454 DNA library dataset) revealed eight different DNA sequence structures (figure 3.10). Structure A, containing two primer sequences (either LTR or *pol*) and no adapters, was the most frequent across all but the 454 DNA library dataset. This result is not unexpected given the reduced ligation efficiency.

Furthermore, we investigated the presence of chimeric sequences which may be formed through the ligation of more than one PCR product during the inter-circularization step (figure 3.10 Structures E and F without and with an adapter incorporated respectively). Our data suggest that the formation of presumptive chimeric DNA products is a rare occurrence, where two out of the three datasets (non-adapter long and the 454 DNA library dataset) without an incorporated adapter (Structure E – figure 3.10) contained 1 percent chimeric sequences, whereas the non-adapter short dataset contained a maximum of 11 percent chimeric sequences. In contrast, chimeric sequences with an adapter sequence ligated (Structure F – figure 3.10) were only identified in 5 percent of sequences within the 454 DNA library dataset. As a final analysis of molecular structure, we also determined the occurrence of single primer amplification of our target regions across our datasets. This was characterized by identification of sequences with a single primer, both without and with a ligated adaptor sequence (figure 3.10 Structures G and H respectively). Linear products were less common compared to inverted sequences with two primer sequences, with the highest percentage of linear products found in both the non-adapter long and non-adapter short datasets (18%). Overall, the sum of linear sequences without and with an adapter sequence (figure 3.10 Structures G and H) across the four datasets was between 14-18 percent. The results indicate that circularization of the fragmented DNA and subsequent iPCR was efficient enabling the preferential (exponential) amplification of circularized DNA versus (non-exponential) linear DNA.

3.3 Discussion

Starting from a limited amount of known sequence to identifying the sequences flanking it is a challenge relevant for numerous applications. Identifying viral integration sites across a host genome is crucial for understanding the integration preferences of viruses and therapeutic viral vectors and to study the biological effects of these integrations. While several molecular techniques exist to study these processes using short read HTS platforms [149, 101, 150, 151, 102, 130, 132], an adaptation of long read iPCR holds several benefits over current methods and has not been explored comprehensively. Sonication-based fragmentation (step 1) enables the random cleavage of DNA across a genome and therefore does not bias the recovery of integration sites in the way that using restriction site digestion does. It is also flexible by allowing optimization of DNA fragment size generation.

In contrast, random fragmentation complicates breakpoint analysis as the ends of the DNA fragments are challenging to identify following circularization. The incorporation of an adaptor sequence on either end of the fragmented DNA (step 2) was designed to abate this issue. This would theoretically aid in the biological interpretation of the sheared DNA breakpoints and the restructuring of inverted sequence reads. It is unclear why ligation efficiency was so low across the standard SiP datasets. However, adapter ligation efficiency has been shown to vary significantly across different library preparation

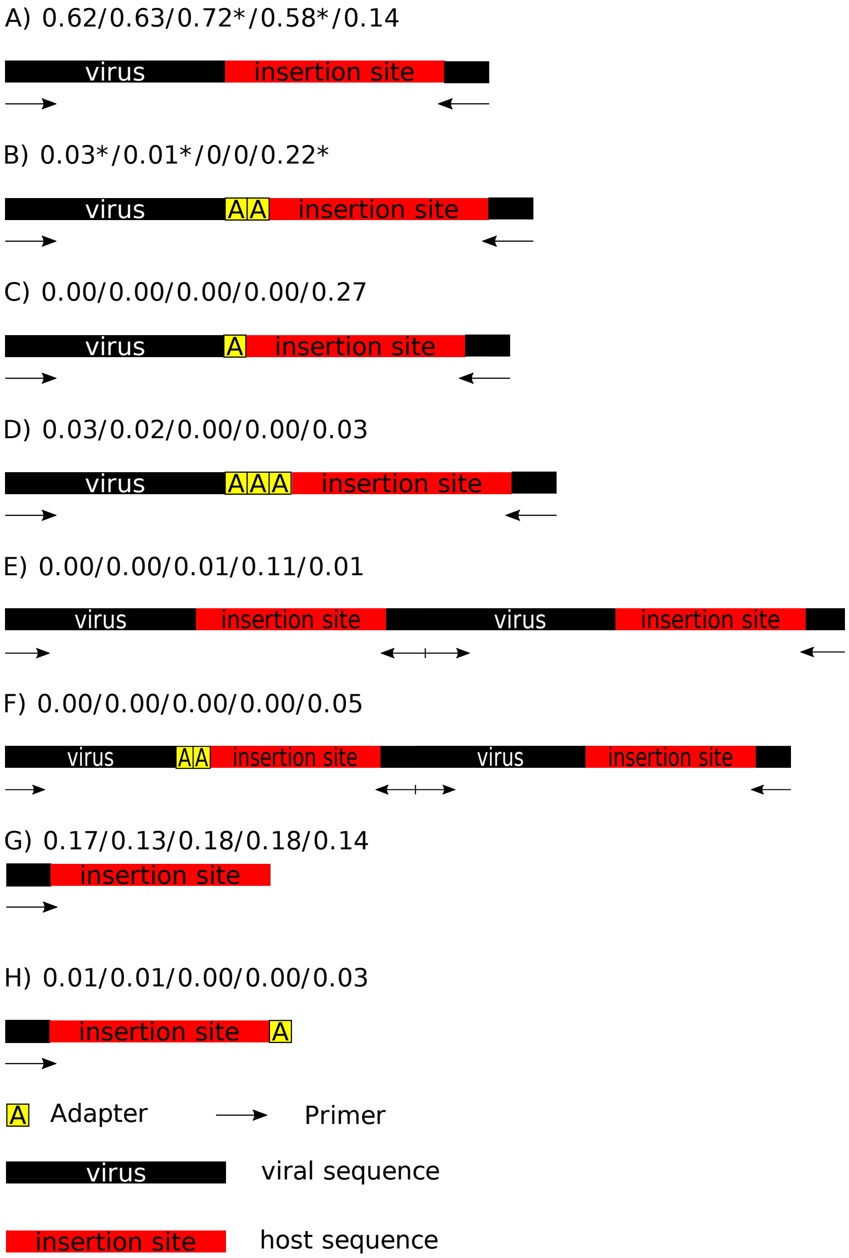


Figure 3.10: Identified Structural Variants of SiP Sequences [3]

Structural sequence variants identified in five SiP datasets are shown (structures A to H). The percentage of reads, where 1.00 = 100%, from each structure and for each dataset is indicated from left to right (adapter long, adapter short, non-adapter long, non-adapter short and the blunt end 454 DNA library datasets). Expected structures were tagged with an asterisk, which were structure A) for the non-adapter long and non-adapter short experiments and B) for adapter short, adapter long and blunt-end 454 DNA library (Illumina adapter) ligation experiments. The displayed structures represent 87 percent of the structures observed for the adapter long dataset and respectively 81 percent for adapter short, 91 percent for non-adapter long, 88 percent for non-adapter short and 89 percent for blunt-end 454 DNA library (Illumina adapter) dataset.

methods. A recent study reported eight out of nine commercially available library kits had a maximum adapter ligation efficiency no greater than 30 percent [169]. Our result on blunt-end ligation efficiency, including those used for generating a 454 DNA library, further exemplifies the limitations of these processes. The blunt-end 454 DNA library method (with the Meyer & Kircher library modifications) [161, 165] is effective, inexpensive and commonly used across genomics studies. However, the reduced adapter ligation efficiency and the halving of useable molecules due to non-directional ligation of identical adapters likely reduces the complexity of sequenced DNA libraries. This is particularly important in studies that use degraded or low amounts of template material (i.e., ancient DNA or museum specimens) as the produced sequence data may not reflect the true DNA diversity of a biological sample. Future studies should examine and compare other adapter ligation techniques such as those used in ssDNA library construction [170, 166] before performing the inter-circularization and amplification steps in SiP.

It is not clear how efficient the circularization process is when using SiP, and like adapter ligation processes, it is possible that the observed diversity is not a true reflection of the diversity in the sample. The analysis of SiP structure variants (figure 3.10) indicates that the vast majority of the reads across the four datasets (82-86%) was inverted. While this suggests that the majority of the DNA has been circularized, the exponential nature of PCR may have masked the non-circularized DNA template, as well as the (less efficient) amplification of linear DNA. Another important consideration is the low amounts of DNA that are suggested in circularization protocols and the subsequent effects this may have on rare variant detection. To minimize these effects, our experimental workflow incorporated several circularization replicates to reduce any potential biases and to maximize recovery of integration sites. Notwithstanding, we recently identified an increased number of KoRV and its recombinants (recKoRV) integration sites [2] in our data compared to those characterized across the koala genome [33, 146]. This suggests that saturation via viral integration site recovery was likely reached across our datasets. However, integrations that occur in few cells (exogenous retroviruses) may require deeper sequence depth to identify.

As a PCR-based method, the effectiveness of SiP will be limited by both the variability and the frequency of the target sequence being amplified. The inverse orientation of the PCR primers designed based on the LTR of a provirus enables concurrent retrieval of 5' and 3' integration sites. Unlike standard PCR methods employed to study viral integration sites, in SiP, both forward and reverse primers can be anchored in close proximity of each other, thereby eliminating the need to prime the reaction inwards from a ligated adapter. Our experiments indicated a total KoRV enrichment rate between 63 and 95 percent for our four standard SiP datasets. Given that these libraries were built from the same sample it is unlikely that the observed differences reflect the true variability across the PCR assays, but are rather due to the varied workflow throughout the experiment.

In the same context, the library generation process employed for PacBio sequencing in these experiments limited the mean size of our DNA libraries, and consequently, the length of the obtainable sequences. While there is likely an upper limit to the size of the circularized products due to steric influences, the analysis of the SiP sequences revealed that fragments of up to 10 kb were successfully circularized, amplified and sequenced. This suggests with the development of longer read platforms (e.g. PacBio Sequel System and Oxford Nanopore MinION) even larger fragments could be enriched and high-throughput sequenced in the future.

While the repeated LTR region further complicates assembling the proviral structure from short reads, SiP introduces its own unique challenges. Unfortunately, given the adapter ligation process was inefficient, restructuring the rearranged iPCR sequences proved challenging. Various algorithms were tested to create consensus sequences for every insertion site. However, most attempts to create a consensus sequence were based on the removal of viral regions followed by a Markov-based clustering algorithm (mcl version 12-135) [96, 97] to compute clusters of high similarity genomic regions. None of these approaches produced consistent results due to duplication of LTRs, low complexity regions within the insertion sites and various unknown structural rearrangements likely due to viral recombination. These various challenges resulted left us unable to assemble complete provirus integrants associated with specific integration sites. However, one major benefit of the coupling of SiP with long-read sequencing was that the majority of integration sites were linked to either *gag* or *env* genes of the provirus. Through a bioinformatics perspective, this simplified re-orientation of the reads compared to data from short-read sequencers.

The ligation of an adapter for structural interpretation of SiP sequences remains an important consideration for both breakpoint identification and the assembly of specific proviral integrants. Importantly, adapter ligation is not a technique that solely affects SiP as it is used across most molecular methods. Despite these challenges, our experiments demonstrate that SiP is a simple, robust and efficient methodology for the analysis of proviral integration sites. The methodology is also highly specific and can be used to characterize any unknown sequence flanking a known sequence; including transgenes and transposable elements for studying genetically modified organisms or host-pathogen interactions. SiP can also be used across taxonomic projects to characterize highly diverse sequences such as the Major Histocompatibility Complex or large genomically variable viruses from short stretches of known sequence. It can also be used in re-sequencing projects to target poorly characterized areas across genomes. While the human genome is arguably the most complete mammalian reference genome, previous assemblies have been found to contain numerous large gaps [171, 172]. Furthermore, structural variation such as indels, duplications, inversions, and tandem repeats remain poorly understood. Several of these can cause a range of Mendelian diseases and can be resolved using SiP or other targeted long-read sequencing applications [173, 174]. SiP is therefore expected to assist broadly across a range of genomic studies.

3.3.1 Transposition of Inverse Polymerase Chain Reaction Products

A secondary objective of this study was to develop a software tool to process SiP reads resulting in reads representing the original biological structure. As the koala genome was not published at the early stages of this project, I attempted to develop host genome reference-free approaches to automatically detect retroviral integration sites.

In the section “SiP Structure Variations” the structural variants found in data retrieved from SiP were described. As indicated in figure 3.11 the structures found were different from the structure expected. Based on the expected read structure I tried to align the adapters attached to DNA fragments before circularization using local pairwise sequence alignment (EMBOSS Smith-Waterman). The strategy was to identify the original sequence ends by adapters, break up the DNA sequence at that point, remove the adapter sequences and in silico ligate the products into the original biological structure.

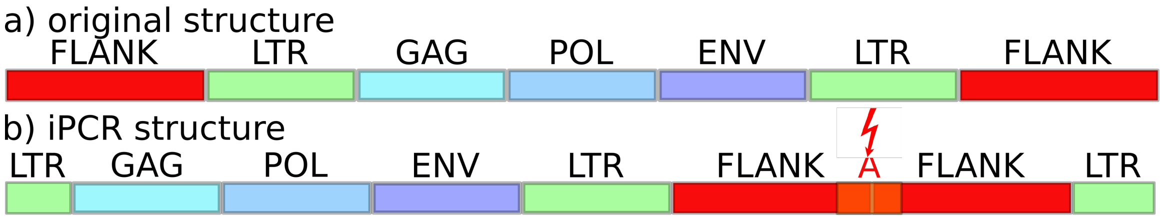


Figure 3.11: Expected Structure of Inverse PCR Reads

As indicated, the circularized sequences have a different order than their biological blueprint. Ideally the product would span the provirus (LTR - green, GAG - light blue, POL - blue, ENV - purple) and cover both integration sites (red "flanks"). Two internal adapter (A) sequences tag the artificial breakpoint (lightning) to open the synthetic circle and restructure the read.

This approach failed due to an adapter ligation efficiency of 2 to 4 percent.

Another approach to profile KoRV integration sites was a combined iterative method using sensitive database searches (BLASTn) for KoRV and PhER sequences, merge these regions on different thresholds ($d = 20; 50; 100\ bp$) extracting homologous regions and repeat that procedure until neither KoRV nor PhER could be detected anymore. The procedure was implemented in Perl. Sequences were filtered for read length, resulting in a file containing only insertion site sequences of KoRV and recKoRV longer than 20 bp. Analogous to the procedure described in chapter 2, these sequences were clustered performing an all-versus-all database search using BLAST. The alignment results were used to perform clustering with tribeMCL. Each cluster was aligned using MAFFT. Consensus sequences were computed from multiple sequence alignments based on 80 percent nucleotide identity. In parallel, cd-hit was used to extract the longest representative sequence from each cluster.

On account of the fact that the koala genome contains unknown structural rearrangements of KoRV and the high number of recombination evens, it is hardly surprising that none of the before-mentioned approaches produced valid results. This concurs well with the challenges of validating specific integration sites using Sanger sequencing shown in supplementary figures A.1-A.11.

3.3.2 Assembling Complete Proviruses at Genomic Loci

As three regions of KoRV were targeted using SiP (5' and 3' LTR and *pol* gene), we attempted to assemble complete proviruses. CCS clusters based on sequence similarity of gDNA were aligned separately using MAFFT (v7.305b) [139], and consensus sequences were produced using BioPerl (Bio::AlignIO) [175]. This attempt at restructuring the inverse reads was not successful since resulting alignments had significant gaps, despite the testing of different algorithms and parameters of MAFFT. To test a multiple alignment algorithm for circular sequences, MARS [88] was assessed based on the same pre-processed clusters, which did not produce plausible consensus sequences. Clustered and unclustered CCS were assembled using Canu (v0.0) [176]. The clustered sequences were assembled once without pre-processing and once with masked repetitive regions (RepeatMasker version open-4.0.7) [177]. A last attempt to restructure the sequences used two separate approaches including, 1) the insilico normalization of the reads (bbnorm) and assembly using Canu and b) splitting reads of a cluster into k-mers (bootstrapped with various k-mer length) and an attempt to assemble them using Velvet (1.2.10) [178, 179]. In figure 3.12 the results in silico fragmented reads in contrast to native SiP products mapped to one of the KoRV integration sites shared with the reference koala genome [146] is shown.

Reads were mapped to the KoRV reference [180] using BLASR [92] with default settings and mapping all subreads using soft-clipping, hard-clipping and without clipping. All intermediate results were used to create reference-based assemblies. The resulting consensus sequences were compared, and a variant calling was performed.

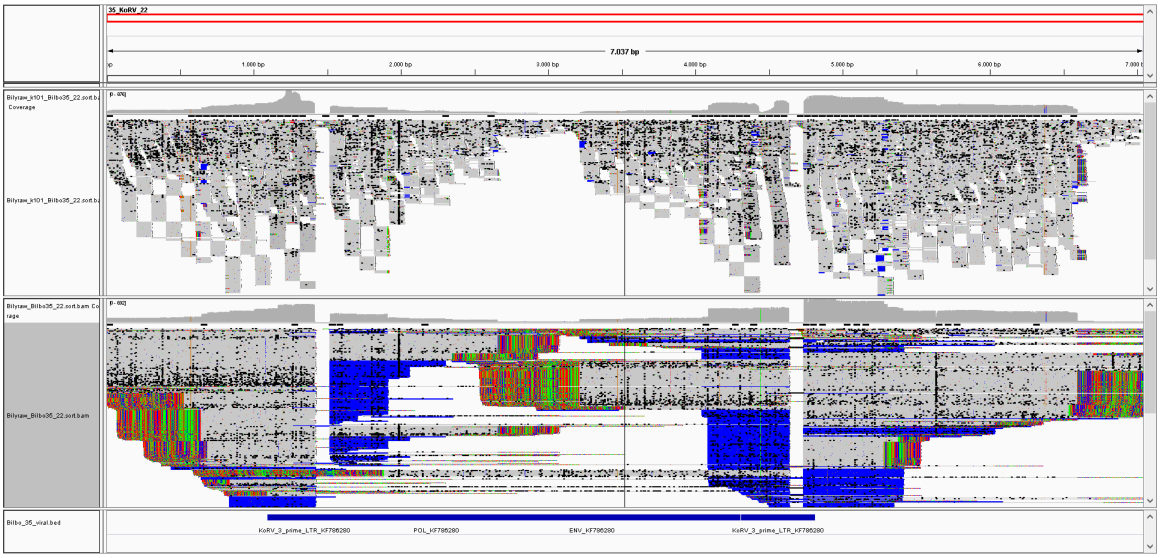

35_KoRV_22
7.037 bp
bp
1.000 bp
2.000 bp
3.000 bp
4.000 bp
5.000 bp
6.000 bp
7.000 bp
Bilyraw_k101_Bilbo35_22.sort.ba
Coverage
[0 - 876]
Bilyraw_k101_Bilbo35_22.sort.ba
Bilyraw_Bilbo35_22.sort.bam Coverage
[0 - 692]
Bilyraw_Bilbo35_22.sort.bam
Bilbo_35_viral.bed
KoRV_3_prime_LTR_KF786280
POL_KF786280
ENV_KF786280
KoRV_3_prime_LTR_KF786280

Figure 3.12: KoRV Integration Site Shared with Koala Reference Genome (KoRV 35)

In the upper panel reads in silico fragmented to a sequence length of 101 bp were mapped to KoRV35 and its genomic flanks. In the bottom panel, the same unprocessed reads are mapped to the locus.

Unfortunately, none of these approaches produced valid results, due to duplication of LTRs, low complexity regions within the insertion sites and various unknown structural rearrangements due to viral recombination and diversity.

Chapter 4

Investigation of Koala Retrovirus in Modern Samples

Endogenous retroviruses (ERVs) are proviral sequences that result from colonization of the host germline by exogenous retroviruses. The majority of ERVs represent defective retroviral copies. However, for most ERVs, endogenization occurred millions of years ago, obscuring the stages by which ERVs become defective and the changes in both virus and host important to the process. The koala retrovirus (KoRV), only recently began invading the germline of the koala (*Phascolarctos cinereus*), permitting analysis of retroviral endogenization on a prospective basis. Here, we report that recombination with host genomic elements disrupts retroviruses during the earliest stages of germ-line invasion. One type of recombinant, designated recKoRV1, was formed by recombination of KoRV with an older degraded retroelement. Many genomic copies of recKoRV1 were detected across koalas. The prevalence of recKoRV1 was higher in northern than in southern Australian koalas, as is the case for KoRV, with differences in recKoRV1 prevalence, but not KoRV prevalence, between inland and coastal New South Wales. At least 15 additional different recombination events between KoRV and the older endogenous retroelement generated distinct recKoRVs with different geographic distributions. All of the identified recombinant viruses appear to have arisen independently and have highly disrupted ORFs, which suggests that recombination with existing degraded endogenous retroelements may be a means by which replication-competent ERVs that enter the germline are degraded.

In humans, about 8 percent of the genome consists of ERV-like elements, comprising a larger proportion of the genome in humans and other species than protein-coding regions within genes [181, 182, 35]. The architecture of the human genome reflects a long evolutionary history of invasions of the germline by infectious retroviruses [181, 182, 183, 19, 184, 185]. Phylogenetic analyses suggest that retroviruses have, from a deep evolutionary perspective, frequently jumped from one species to another moreover, invaded the germlines of new hosts [186, 29, 187, 188, 189]. Almost all known ERVs completed invasion of their host germlines millions of years ago, obscuring the early events critical to the invasion process. An exception is KoRV.

KoRV is a full-length replication-competent endogenous retrovirus, the titer of which is correlated with chlamydiosis and hematopoietic neoplasia [118, 27]. KoRV is thought to spread in koalas both horizontally by infection and vertically as an endogenous genomic element [190, 121]. KoRV has a clinal geographic distribution among koalas; while 100 percent of northern Australian koala populations carry KoRV, the prevalence and copy number of KoRV is significantly reduced in southern Australia [27, 34, 191]. Unlike other ERVs, KoRV is not present in the germline of all members of the host species [119]. Ancient DNA studies have shown that KoRV was ubiquitous in Queensland koalas in the late 19th century [121]. Molecular dating places the initial entry of KoRV into the koala germline within the past 50,000 y [121, 31]. These studies strongly indicate that KoRV, unlike most known vertebrate ERVs, is in the early stages of the endogenization process in its koala host [34, 119].

In vertebrates, many ERVs are found at fixed positions in the genome across all members of the host species. By contrast, there is substantial variation in the host genomic integration sites for endogenous KoRV across koalas, with a very high degree of insertional polymorphism. Among modern and museum samples of koalas in Queensland, the population with the highest abundance of KoRV, only a small proportion of KoRV integrations are shared among unrelated koalas [31, 1, 123]. The lack of fixed KoRV proviral insertions among koalas and the low proportion of shared integration sites among koalas that carry endogenous KoRV provide further evidence for a recent invasion of the koala germline by KoRV.

Most endogenous retroviruses in other species have highly disrupted proviral genomes, and those present in higher copy numbers across the germline often show deletion of the proviral envelope gene (*env*), which codes for the viral envelope. ERVs that lack *env* have been found to be "superspreaders" i.e., elements that have reached a high copy number within the host germline [192]. Through recombination, exogenous retroviruses can exchange genetic information with ERVs in infected individuals. For example, murine leukemia virus (MLV) can recombine with endogenous MLVs to generate novel viruses, in effect remobilizing part of the ERV sequences [193]. Recombination may also render ERVs defective, e.g. through disruption of ORFs or the generation of solo long terminal repeats (LTRs) [194]. However, the role of recombination during the early stages of retroviral genomic invasion has not been directly examined. Here, we provide evidence that older endogenous retroelements recombine with and degrade invading retroviral genomes, even when the homology between them is limited. This occurs early during the retroviral invasion and disrupts the invading retrovirus while simultaneously remobilizing an existing retroelement recombination partner within the host genome. By disrupting retroviruses invading the germline, the process likely accelerates the retroviral transition from horizontal to vertical transmission, which is expected to benefit the host species.

4.1 Methods

As described in chapter 3, we developed a method to examine retroviral integration sites to investigate the endogenization of KoRV in a modern zoo koala taking advantage of third-generation sequencing. Following the protocol of sonication iPCR, we extracted KoRV integration sites from spleen tissue of koala "Bilyarra" (Pci-SN241), which died in July 2014. "Bilyarra" was a 16 years old male from Tiergarten Schönbrunn in Vienna (Austria).

As a reference, we used the koala genome retrieved from wild (southeast Queensland, Upper Brookfield), female koala "Bilbo" (Australian Museum registration M.47724), sampled following euthanasia due to severe chlamydiosis (20 August 2015) [33, 146]. In addition, we used sequences from 175 different koalas from four different studies for LTR comparison (see table 4.1).

4.1.1 Koala Samples, PCR, and Sequencing

Four sources of genomic data were employed in the current study. Table 4.1 lists the koalas sampled. It indicates for each set of koalas the sequencing platform used to generate datasets, the source of the data (generated for the current study or mined from existing data), and also summarizes the results obtained by analyzing new datasets or reanalyzing previously generated datasets.

Short-Read Illumina Sequences

We used Illumina-based genome sequences (unassembled) from two koalas, "Pacific Chocolate" and "Birke" (table 4.1) [146]. This dataset was screened for KoRV and PhER breakpoints. Additionally, existing Illumina datasets were reexamined for KoRV and PhER breakpoints (table 4.1) [123]. "Pacific Chocolate" (a wild-born New South Wales koala) and "Birke" (a wild-born Queensland koala) had been sequenced with 100x Illumina short-read coverage [146]. Mirali (PCI-SN265) was a zoo koala (northern Australian lineage) from the Vienna Zoo, Tierpark Schönbrunn. It had been Illumina sequenced after KoRV enriched hybridization capture. The dataset is described in reference [123]. Archived museum samples of six koalas collected in Queensland between 1870 and 1938 had been Illumina sequenced after KoRV enriched hybridization capture, as described in reference [123]. All of these koala datasets were examined for the presence or absence of the recKoRV recombination breakpoints shown in figure 4.2, with results by koala shown in figure 4.8 and table 4.4.

PacBio Long-Read Sequencing

Two koalas were sequenced using the PacBio platform: "Bilbo" (a wild koala from Upper Brookfield, Queensland) and "Bilyarra" (from the Tierpark Schönbrunn, Vienna) (table 4.1).

"Bilbo" is the koala for which the koala reference genome has recently been described [146]. Briefly, the long-read genome assembly used in this work is version: phaCin_unsw_v4.1 deposited in DDBJ/ENA/GenBank under the accession GCA_002099425.1 with the genome assembly project registered under BioProject PRJNA359763. High molecular weight (HMW) DNA was extracted from female koala spleen (Australian Museum registration M.47724) using Genomic-Tip 100/G columns (Qiagen) and DNA Buffer set (Qiagen). Fifteen SMRTbell libraries were prepared and sequenced on the Pacific Biosciences RS II platform with a total of 272 SMRT cells sequenced to give an estimated overall coverage of 57.3X based on a genome size of 3.5 Gbp. After filtering low-quality and duplicate reads, approximately 57.3-fold read coverage was used for assembly. Primary contigs (homozygous regions) made up 3.19 Gbp of the assembled genome, comprising 1,906 contigs, with an N50 of 11.6 Mbp and sizes ranging up to 40.6 Mbp. An assembly produced using Falcon (v.0.3.0) including the 5225 alternate contigs of heterozygous regions yielded a 3.42 Gbp assembly with an N50 of 48.8 kbp. Approximately 30-fold coverage of Illumina short reads were used to polish the assembly with Pilon [146].

Spleen tissue of "Bilyarra" was used to extract DNA for the current study using the QIAamp DNA Minikit (Qiagen). The integration sites and KoRV and recKoRV sequences were enriched prior to PacBio sequencing as described below using iPCR and PacBio sequencing. The data generated from these koalas were of specific relevance to defining full length recKoRV1 in figure 4.2 and integration sites and LTRs described in figures 4.8 and 4.9 which could not be accomplished for individual loci using Illumina data.

PCR-Based Screening of recKoRV1 from 166 Koala DNA Samples

A total of 166 wild koala DNA samples was collected by J. Meers, P. Young and their associates [34]. DNA was extracted from these 166 samples using the Blood & Tissue DNA Extraction Kit (Qiagen) or was provided by collaborators. The DNA was amplified for the recKoRV 3' breakpoint (with the koala actin gene used as a positive control for DNA quality). The amplified recKoRV PCR products were Sanger sequenced to establish their identity as the recKoRV1 3' breakpoint. PCR screening of recKoRV1 from the 166 koalas (figure 4.10) involved two primer sets

Ya_recKorRV1-F	5'- GCT GCT TGA TTT GGA TGT GA -3'
Ya_recKoRV1-R	5'- GAG GAG TAG CAG GGG ACC AG -3'
recKoRV-F1	5'- TGT GAA TAT CCC TGG CAG CCG CG -3'
KoR27-R	5'- GAG TAA CAG AAG GAG GAG TAG CAG -3'.

Sequences were trimmed and visualized using the alignment and assembly program Vector NTI advance 11 (Invitrogen). It should be noted that using this PCR strategy, recKoRV1 and recKoRV2 cannot be readily distinguished [33]. However, recKoRV2 was generally rare in both the koala reference genome and in Pci-SN241, and the two recombinants likely are very closely related so this would not change the interpretation of the results. The resulting data is presented in figure 4.10.

4.1.2 Inverse PCR and PacBio Sequencing of "Bilyarra" to Determine KoRV and recKoRV Sequences

gDNA was extracted, fragmented and blunt-ended as previously described (see chapter 3). Briefly, DNA was extracted using a standard silica-based tissue extraction kit, the QIAamp DNA Minikit, (Qiagen, Hilden Germany). DNA was then fragmented using a Covaris ultrasonicator, which produced an average DNA fragment size of 2-7 kbp in length. Sheared DNA was subsequently blunt-end repaired using the commercially available Fast DNA End Repair kit (Thermo Scientific) in triplicate [3].

To find the optimal ligation conditions for subsequent iPCR, we performed a series of nine ligations using a gradient of (total) input blunt-ended DNA. Briefly, ligation reactions were set up using a commercially available T4 DNA Ligase kit (5 U/μL) (Thermo Scientific). Ligation was performed in a thermal cycler at 16 °C for 16 hours followed by enzyme inactivation at 70 °C for 5 min. Given the minuscule starting DNA amounts, all ligations were performed in triplicate [3].

Inverse PCR was performed as previously described (chapter 3) using primers to the KoRV LTR. A primer set was designed using Primer3Plus software [163] targeting a conserved region in the middle of the KoRV LTR (iPCR_LTR_F: TGCATCCGGAGTTGTGTTCG; iPCR_LTR_R: AAAAGCGCGGGTACAGAAGC). The optimal circularization PCR product in each gradient was chosen by the analysis of three criteria on the TapeStation. This included (i) considering the DNA amount per micro-liter of iPCR product in the 2-7 kbp range, (ii) the average length distribution between a 600 bp-7 kbp range,

Table 4.1: Koala Samples and Datasets Utilized [2]

Koala	**Wild/zoo***	**Sample sources**	**Sequence type**	**Source**
Bilyarra	SN241	Vienna Zoo (Tierpark Schönbrunn)	SiP PacBio	This paper
Bilbo	Wild	Australian Museum registration M.47724, Upper Brookfield Queensland	PacBio genome assembly	[33, 146]
Pacific Chocolate	Wild	Port Macquarie, New South Wales	Illumina sequences	[33, 122]
Birke	Wild	Australian Zoo Wildlife Hospital in Queensland	Illumina sequences	[33, 122]
One zoo and 6 museum specimens	SN265 and historical/wild	Vienna Zoo (Tierpark Schönbrunn) and various museums	Hybridization capture Illumina sequences	[123]
Samples of 166 koalas	Wild	Collected across koala range in Australia	PCR and Sanger sequencing	[34]

** SN indicates the European koala studbook number for samples from zoological collections.*

The 166 wild koalas in reference [34] were sampled from across their geographic range. "Pacific Chocolate" was from New South Wales. All other samples were derived from the Queensland koala population, including all zoo koalas and museum specimens. Database refers to National Center for Biotechnology Information GenBank and Sequence Read Archive.

(iii) and the percentage of DNA within the 2-7 kbp range. Following these criteria, a 40 ng input DNA (conc. 0.8 ng/μL in circularization) were used.

PCR products were submitted for PacBio library construction and sequencing to the Max Delbrück Center, Berlin. PCR products were purified using AMPure XP beads (Beckman Coulter), first at a concentration of 0.4X followed by a subsequent purification of the supernatant at 0.6X. The resulting four samples were prepared as sequencing libraries using the PacBio (Pacific Biosciences, Menlo Park, CA) 5 kb template prep protocol and the SMRTbell^{TM} Template Prep Kit 1.0 following the manufacturer's recommended protocols. Library concentration and fragment length were verified using the Qubit 2.0 fluorometer (Life Technologies) and the 2100 Agilent Bioanalyzer, using the 12000 DNA chemistry (Agilent Technologies). The estimated average lengths for the short and large insert libraries were 1600 bp and 3500 bp, respectively. Sequencing on the PacBio RSII platform used the MagBead Standard protocol, C4 chemistry and P6 polymerase on a single v3 Single-Molecule Real-Time (SMRT) cell with 1×180 min movie for each library (a total of 4 libraries). The reads from the insert sequence were processed within the SMRTPortal browser (minimum full pass = 1; and a minimum predicted accuracy of 90).

Amplification from integration site to integration site for 11 loci identified four of the loci as being different recKoRVs i.e., they had recombination breakpoints that differed from recKoRV1 (table 4.1.3). Most sequences that turned out to be recKoRVs other than recKoRV1 mapped relatively poorly initially to recKoRV1 whereas those that were confirmed mapped well. Sequencing of the 5' breakpoint was particularly challenging due to large numbers of homopolymer stretches, and several products could not be sequenced. Three loci could not be amplified, likely due to the low complexity sequences flanking the integrations and the difficulty of amplifying a 6.4 kbp product in the presence of the empty site on the opposing chromosome. After of mapping and Sanger sequencing, of 14 integrants putatively identified as recKoRV1, and ten were confirmed to be recKoRV1.

4.1.3 Bioinformatics Analysis

Analyses of PacBio Sequences to Identify Integration Sites and Determine Whether the Integrants were KoRV or recKoRV Proviruses

To isolate the host genomic sequences flanking integration sites for KoRV and recKoRV, the KoRV containing reads were aligned to the KoRV-A or -B reference sequences (AB721500.1; KC779547) using BLASTn. Regions homologous to the reference sequences were removed. The isolated host genomic sequences flanking integrations sites were clustered using Tribe-MCL (I=1.4), a Markov cluster-based approach, processing distance-based information of a BLASTn matrix for all KoRV containing reads [96]. The recKoRV1 containing reads were aligned using BLASTn to the KoRV-A and -B reference sequences, as well as to PhER, all known recKoRV breakpoints and the consensus sequence of recKoRV1. Regions homologous to any of the reference sequences were removed. The isolated flanking regions were clustered using Tribe-MCL (I=4). A consensus sequence for every cluster was created by constructing a multiple sequence alignment using MAFFT (v7.305b) [77] and computing a consensus sequence using the Perl module BioPerl::SimpleAlign (30% identity, gap removal) [175]. Raw "Bilyarra" circular consensus sequences (ccs), KoRV and recKoRV1 insertion site flanking sequences(a consensus of all sequences) were mapped to the assembled genome of koala "Bilbo" using the Burrows-Wheeler alignment (BWA BWA-SW default) for long sequences [90]. Regions of interest

Table 4.2: recKoRV1 Classification Summary

	bp12			bp17			rec1
Insertion	iPCR 99% 101bp	Sanger	bwa	iPCR 99% 101bp	Sanger	bwa	status
Scaf00003	16	didn't amplify	✓	107	didn't amplify	✓	(✓)
Scaf00014	0	X	X	1	(X)	X	-*
Scaf00021	4	(✓)	✓	37	✓	✓	✓
Scaf00024	0 (3;90%)	✓	(✓)	39	✓	(✓)	✓
Scaf00037_16	6	X	✓	65	✓	✓	(✓)
Scaf00037_17	22	X	✓	224	✓	✓	(✓)
Scaf00069	25	X	✓	124	✓	✓	(✓)
Scaf00079	8	X	✓	46	✓	✓	(✓)
Scaf00083	15	(✓)	(✓)	35	✓	✓	✓
Scaf00096	12	X	X	168	X	✓	-
Scaf00164	14	didn't amplify	✓	136	didn't amplify	✓	(✓)
Scaf00173	21	X	✓	2	X	(X)	-
Scaf00234	1	didn't amplify	(X)	184	didn't amplify	✓	-
Scaf00275	5	X	✓	55	✓	✓	✓**

** similarities to bp17 present, polymerase gene (pol), PhER present was a former heterozygous candidate*

*** 275 dummy completely covered*

Modified from [2]

were determined using Bedtools [195]. We examined regions covered by at least 30 ccs in "Bilyarra", and the regions of interest were manually annotated. Each recKoRV integration site determined using the described bioinformatics approaches was then confirmed by PCR including primers based on the regions flanking integration sites, using Sanger sequencing to determine whether the elements and structures identified were consistent with the bioinformatics analysis of the iPCR products (figures 4.1, A.4, A.5, table 4.1.3, supplementary figures A.1 - A.11).

Bioinformatics Analysis of Koala Illumina Sequence Data

Collaborators performed the following analysis. Hybridization capture of KoRV is described in [123]. The recKoRV1 breakpoint sequences in "Pacific Chocolate" and "Birke" were initially detected in transcriptome sequences [122]. Subsequently, Illumina 100 bp genomic sequence libraries from both of these koalas [146] were screened. First, a subset of reads enriched in KoRV sequences was produced using the fastmap mode of bwa to align reads to a reference KoRV genome sequence. Second, using BLASTn the KoRV-enriched set of reads was filtered to remove reads with full-length alignments to KoRV, leaving reads of potentially chimeric sequences. Finally, BLASTn was used to query these potential chimeric reads to a PhER sequence. Next-generation sequence data from archival samples, obtained from [123], were filtered using cutadapt v1.8.1 for adaptor sequence, low quality reads, and fragments shorter than 30 bp [134]. Sequence data that passed the quality filters were aligned to breakpoints identified in recKoRV1 using BWA version 0.7.15-r1140 and the mem algorithm with default settings [196]. Aligned data were further processed using samtools for clonal read removal. Identified breakpoints were confirmed visually using Geneious 7.1 [162]. The results of the breakpoint characterization are shown in figure 4.2, figure 4.8 and table 4.4.

Koala Transcriptome Analyses

Expression of PhER was tested by searching for PhER sequences in the transcriptomes reported by Hobbs et al. (2014) [122] . An 8 kpb PhER sequence (Hobbs et al., 2017) [33] was used as a query in BLASTn searches of "Pacific Chocolate" and "Birke" transcriptome databases. In both cases , the searches revealed PhER transcripts distinct from those forming part of recKoRV, and which notably included those parts of PhER that are not incorporated into recKoRV.

Network Analyses of LTRs

Both "Bilbo" and "Bilyarra" had KoRV and recKoRV integrations with the 5' and 3' LTRs belonging to different LTR groups suggesting that gene conversion or recombination, both observed in other proviruses, had occurred, precluding accurate dating to find individual integrations. PacBio generated KoRV and recKoRV LTRs from all insertion sites of "Bilbo" and "Bilyarra" were aligned with sequences of LTRs from KoRV integrations in ten different koalas examined in [141], and with KoRV-A (AB721500.1) and KoRV-B (KC779547.1). The iPCR primer gaps were removed in all sequences. Multiple sequence alignment was performed using MAFFT L-INS-i [77]. The alignment was cropped to the most conserved regions (89% identity) on both ends, realigned and manually curated. A haplotype network was constructed using the R [197] package Pegas [198] with the distance model "indelblock", performing an iterative refinement for the smallest sum of distances. The results of the analysis are shown in figure 4.9. Bioinformatics analysis of koala Illumina sequence data hybridization capture of KoRV is described in Tsangaras et al. 2014 [123]. The recKoRV1 breakpoint sequences in "Pacific Chocolate" and "Birke" were initially detected in transcriptome sequences [33]. Subsequently, Illumina 100 bp genomic sequence libraries from both of these koalas [146] were screened. First, a subset of reads enriched in KoRV sequences was produced using the fastmap mode of bwa to align reads to a reference KoRV genome sequence. Second, using BLASTn the KoRV-enriched set of reads was filtered to remove reads with full-length alignments to KoRV, leaving reads of potentially chimeric sequences. Finally, BLASTn was used to query these potential chimeric reads to the PhER sequence. Next-generation sequence data from archival samples, obtained from [123], were filtered using cutadapt v1.8.1 for adaptor sequence, low quality reads, and fragments shorter than 30 bp [134]. Sequence data that passed the quality filters were aligned to breakpoints identified in recKoRV1 using BWA version 0.7.15-r1140 and the mem algorithm with default settings [196]. Aligned data were further processed using samtools for clonal read removal. Identified breakpoints were confirmed visually using Geneious 7.1 [162].The results of the breakpoint characterization are shown in figure 4.2, figure 4.8 and table 4.4. Koala transcriptome analysis Expression of PhER was tested by searching for PhER sequences in the transcriptomes reported by Hobbs et al. (2014)[122]. An 8 kbp PhER sequence (Hobbs et al., 2017) [33] was used as a query in BLASTn searches of "Pacific Chocolate" and "Birke" transcriptome databases. In both cases, the searches revealed PhER transcripts distinct from those forming part of recKoRV, and which notably included those parts of PhER that are not incorporated into recKoRV. The results of this analysis are shown in table 4.4 and supplemental tables A.1,A.2,A.3.

4.2 Results

4.2.1 The Advantages of Long Read Sequence Technology for Retroviral Analysis

KoRV has generally been studied using 454 FLX or Illumina-based short-read sequencing approaches [121, 123]. These approaches have a number of limitations. First, identified polymorphisms in KoRV cannot generally be put in phase with other polymorphisms. Second, only small structural differences among KoRV sequences, such as short indels, can be detected. Large deletions and recombination events are missed given that reads are of short length. Third, the specific host integration site cannot be identified for polymorphisms or KoRV sequences because reads are not long enough to cover the DNA region from the integration site to the KoRV genes. Thus, KoRV variation has been studied in the aggregate by mapping reads to full-length proviruses. Using this method, little variation has been detected [121, 123]. By contrast, the current study used PacBio technology, which produces long sequence reads. The koala genome was sequenced using this technology, and we here sequenced individual KoRV proviruses (exemplarily shown in figure 4.1). This permitted us to identify structural variation across individual KoRV proviruses, link KoRV variants to genomic loci in the koala host and determine the position and copy number for each type of variation detected among KoRV proviruses. A complex evolutionary history was revealed for KoRV.

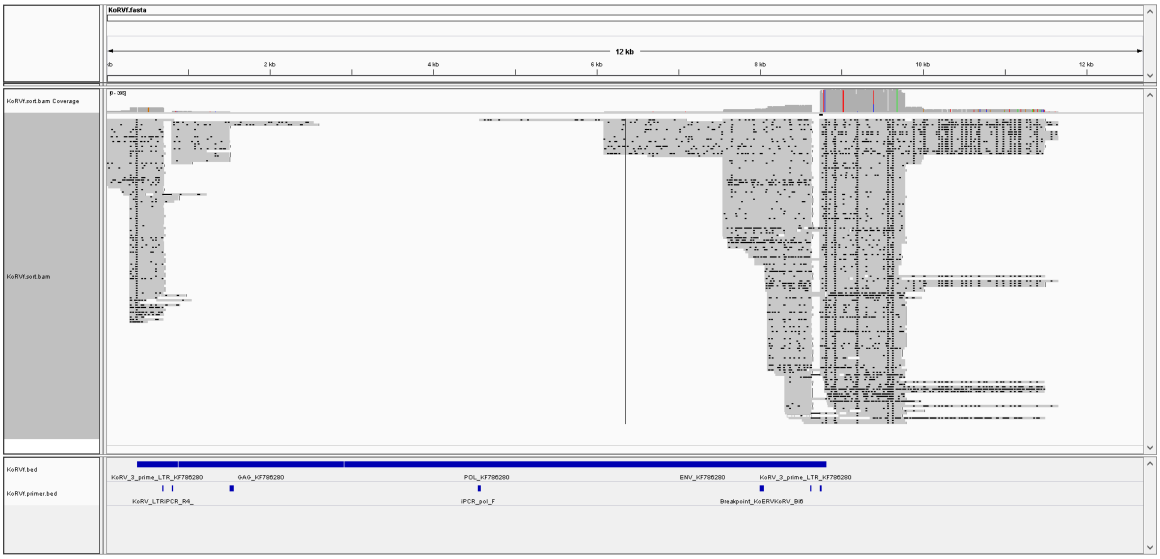


Figure 4.1: Example of Read Mapping from Captured Forward KoRV Integration Site

Mapping of “Bilyarra” sequences assigned to an integration site in scafold 322 of the reference genome (“Bilbo”). The reference is a synthetic construct of a KoRV provirus at the prior determined position of the genome. The blue bar at the bottom shows the different KoRV domains. Primers and recKoRV1 breakpoints are indicated as blue boxes below the bar. The 3’ integration site could be covered by approximately 1 kb of sequence length (insertion close to start position of scaffold). A clear gap could be seen in the LTR region, which is caused by the 99 bp distance between the two LTR primers. At the 3’ site nearly the complete GAG domain was covered, while on the 5’ site the ENV domain was covered. A least one POL primed read could be assigned to this integration site covering parts of the POL domain. The 5’ genomic region was covered by reads longer than 2 kbp.

4.2.2 Discovery of recKoRV1 [2]

PacBio sequencing of one koala (“Bilbo” [146]), the individual used to sequence and assemble the koala genome, has identified a proviral integrant called recombinant koala retrovirus 1 (recKoRV1), which includes the 5’ KoRV LTR followed by the group-specific antigen (GAG) leader region to position 1,177 [33]. It also includes the KoRV region from position 7,619 of the *env* gene including the complete 3’ LTR. However, the sequence between these two fragments of KoRV is derived from another retroelement, designated the Phascolarctos endogenous retroelement (PhER) (figure 4.2) [33]. PhER has partial homology to Repbase [199], but has no intact protein-coding regions except potentially in the *env* region [33, 122]. PhER has been found to be a transcriptionally expressed high copy number ERV (≈30-40 full-length elements and hundreds of solo LTRs or fragmented copies).

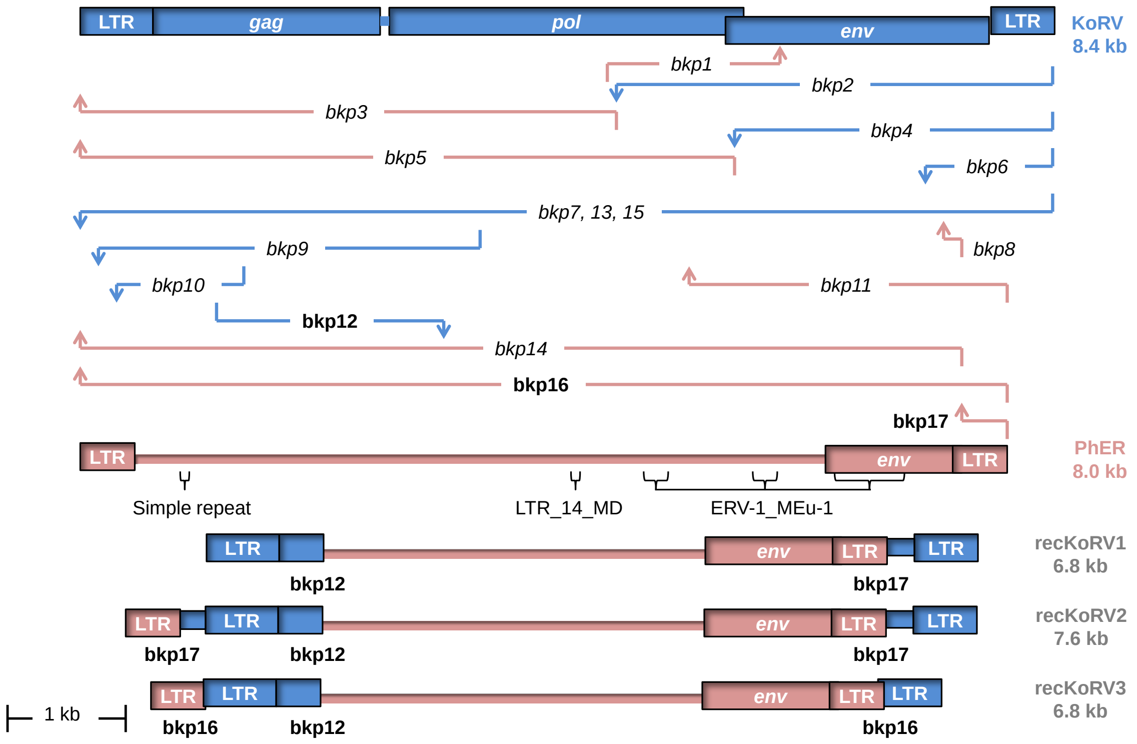


Figure 4.2: Recombination Breakpoints of KoRV and PhER [2]

Breakpoints in KoRV-PhER recombinants. The genomic structures of KoRV (blue) and PhER (pink) [33] are shown, including genes, LTRs, and Repbase repeat motifs identified in PhER. Locations of breakpoints (bkps) in 17 recombinant sequences are represented by arrows, with pink upward-directed arrows used when PhER sequence is 5' of the breakpoint, and blue downward directed arrows when KoRV sequence is 5' of the breakpoint. For bkps within an LTR sequence, only one of the possible alignments is shown. Three recombinant sequences from long read (Pacbio) sequence datasets allowed assignment of breakpoints to recombinant elements recKoRV1, recKoRV2, and recKoRV3. Breakpoints identified only in short-read (Illumina) sequence datasets are italicized.

4.2.3 Other Recombinants Between KoRV and PhER

In the current study, we examined an unrelated koala ("Bilyarra"; table 4.1) to characterize recKoRV. We also identified KoRV-PhER recombination breakpoints and used them as queries to screen existing Illumina sequence datasets that had been previously generated but never examined for KoRV recombinants. Proviral integration sharing among koalas was examined on a per locus basis, while the presence or absence of specific recombination breakpoints was examined in the aggregate (table 4.1 gives details regarding the koalas and datasets). Along with the two recombination sites of recKoRV1, an additional

15 KoRV-PhER recombinant sequences were identified (see figure 4.2 table 4.4). Although KoRV and PhER had dissimilar sequences, at five of the recombination breakpoints, we identified microhomologies, short matching sequences that were shared at a breakpoint by both KoRV and PhER. These microhomologies may have enabled the recombination between the two elements at various breakpoints (see table 4.4) [200, 201, 202]. Of 17 recombination breakpoints identified, all but three were within 1,500 bp of the ends of the KoRV genome (figure 4.2). Most breakpoint sequences were determined using only short Illumina reads, and so it was not possible to determine the structure of recombinants or characterize the integration sites.

A mapping of target enriched "Bilyarra" reads to the 17 recombination breakpoints identified was performed using BWA-SW. No reads were found spanning breakpoints found in "Pacific Chocolates". Four breakpoints from "Birke" could be spanned by reads from the current study. The results are summarized in table 4.3. As shown in figure 4.3 to 4.7 the recKoRV1 breakpoints bp12 and bp17 were covered entirely by sequences retrieved from koala "Bilyarra", including some single nucleotide polymorphisms presumably deriving from different integration sites. Breakpoint bp13 is spanned by reads, but slightly varies from the reference.

Table 4.3: Reads Spanning Breakpoints bp12, bp13, bp16, bp17

	No. of reads spanning BP			
ID of BP	**adapter long**	**adapter short**	**no adapter long**	**no adapter short**
bp12	69	32	65	47
bp13	0	43*	0	0
bp16	1	0	0	2
bp17	29	58	91	192

The number of Illumina sequences among koala datasets mapping to proviral recKoRV1 breakpoints far exceeded those mapping to any of the other recombination breakpoints identified (see figure 4.8 table 4.4) [33, 146]. We, therefore, focused on the evolutionary history of the recKoRV1 subtype of recombinants.

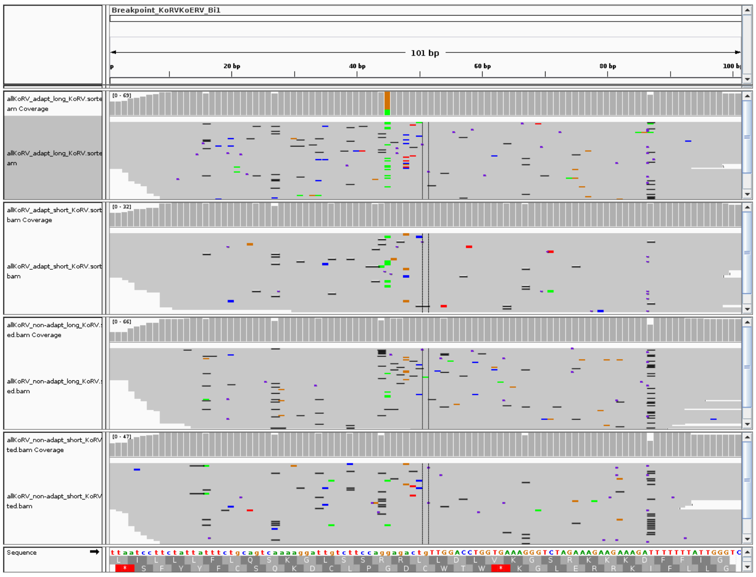

Breakpoint_KoRVKoERV_Bi1
101 bp
20 bp
40 bp
60 bp
80 bp
100 bp
allKoRV_adapt_long_KoRV.sorted.bam Coverage
allKoRV_adapt_long_KoRV.sorted.bam
allKoRV_adapt_short_KoRV.sorted.bam Coverage
allKoRV_adapt_short_KoRV.sorted.bam
allKoRV_non-adapt_long_KoRV.sorted.bam Coverage
allKoRV_non-adapt_long_KoRV.sorted.bam
allKoRV_non-adapt_short_KoRV.sorted.bam Coverage
allKoRV_non-adapt_short_KoRV.sorted.bam
Sequence
[0 - 69]
[0 - 32]
[0 - 66]
[0 - 47]

Figure 4.3: Mapping KoRV Positive Reads to Breakpoint “Birke” bp12

Breakpoint bp12 was covered from read of all experiments. This is expected since it is one of the recKoRV1 breakpoints, which is the most frequent recombination of KoRV and PhER

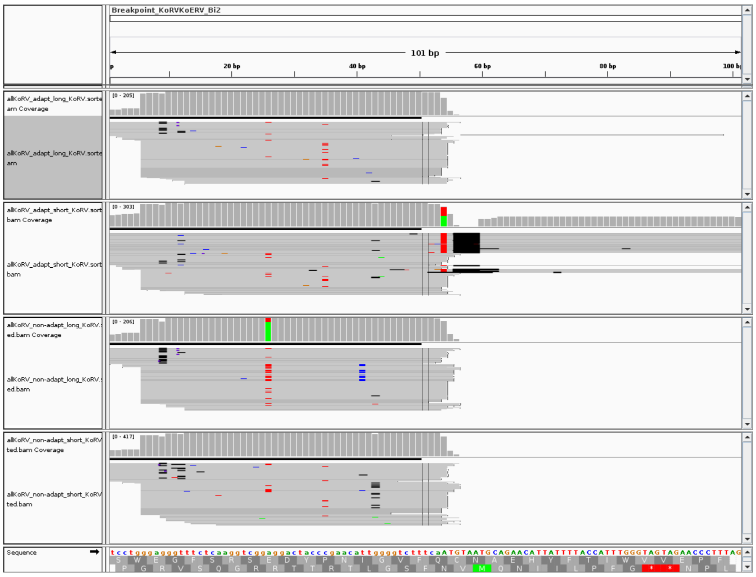

Breakpoint_KoRVKoERV_Bi2
101 bp
20 bp
40 bp
60 bp
80 bp
100 bp
allKoRV_adapt_long_KoRV.sorted.bam Coverage
allKoRV_adapt_long_KoRV.sorted.bam
allKoRV_adapt_short_KoRV.sorted.bam Coverage
allKoRV_adapt_short_KoRV.sorted.bam
allKoRV_non-adapt_long_KoRV.sorted.bam Coverage
allKoRV_non-adapt_long_KoRV.sorted.bam
allKoRV_non-adapt_short_KoRV.sorted.bam Coverage
allKoRV_non-adapt_short_KoRV.sorted.bam
Sequence
[0 - 205]
[0 - 303]
[0 - 206]
[0 - 417]

Figure 4.4: Mapping KoRV Positive Reads to Breakpoint “Birke” bp13

Breakpoint bp13 shows up only in the short-read dataset with adapters. 43 reads span this breakpoint. They differ from the references directly after the breakpoint in the PhER sequence. Our data show a substitution of thymine at the first position instead of adenine, followed by a thymine, which is the same in the reference, followed by a 4-7 bp deletion.

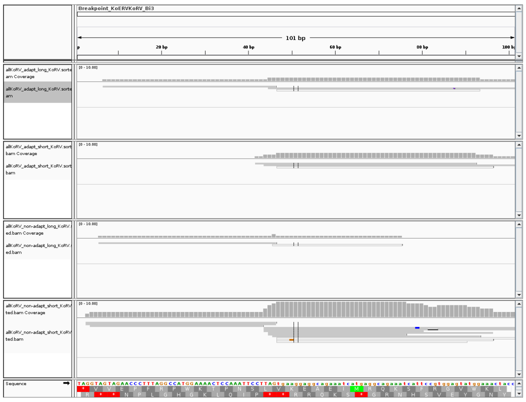


Figure 4.5: Mapping KoRV Positive Reads to Breakpoint “Birke” bp14

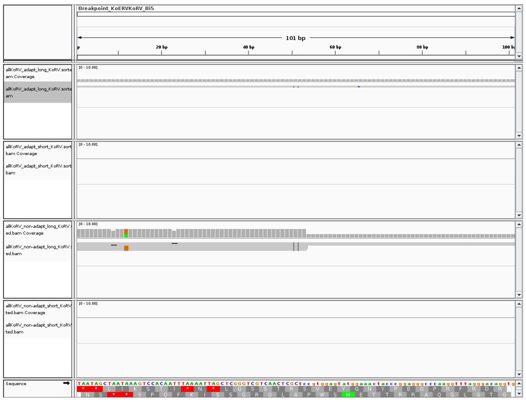


Figure 4.6: Mapping KoRV Positive Reads to Breakpoint “Birke” bp16

For completeness, breakpoints bp14 and bp16 are shown. Bp14 was not spanned by sequences retrieved from “Bilyarra”, whereas Bp16 was only covered by a few reads

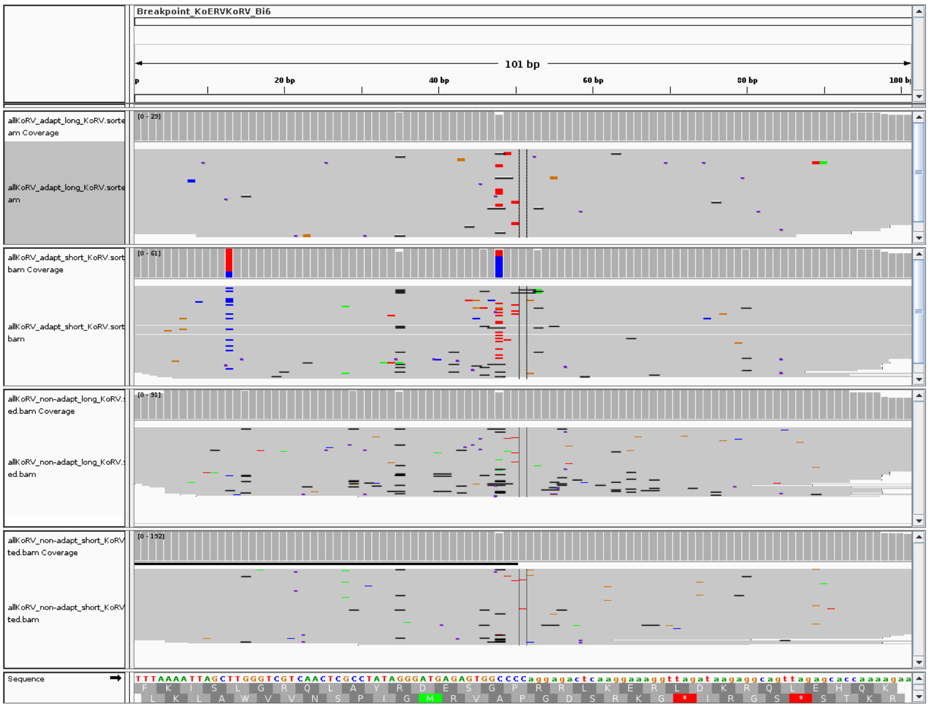

Breakpoint_KoERVKoRV_Bi6
101 bp
20 bp
40 bp
60 bp
80 bp
100 bp
allKoRV_adapt_long_KoRV.sorted.bam Coverage
[0 - 29]
allKoRV_adapt_long_KoRV.sorted.bam
allKoRV_adapt_short_KoRV.sorted.bam Coverage
[0 - 61]
allKoRV_adapt_short_KoRV.sorted.bam
allKoRV_non-adapt_long_KoRV.sorted.bam Coverage
[0 - 91]
allKoRV_non-adapt_long_KoRV.sorted.bam
allKoRV_non-adapt_short_KoRV.sorted.bam Coverage
[0 - 192]
allKoRV_non-adapt_short_KoRV.sorted.bam
Sequence
TTTAAAATTAGCTTGGGTCGTCAACTCGCCTATAGGGATGAGAGTGGCCCCaggagactcaaggaaaggttagataagaggcagttagagcaccaaaagaa

Figure 4.7: Mapping KoRV Positive Reads to Breakpoint “Birke” bp17

Breakpoint bp17 was covered from read of all experiments. This is expected since it is one of the recKoRV1 breakpoints, which is the most frequent recombination of KoRV and PhER

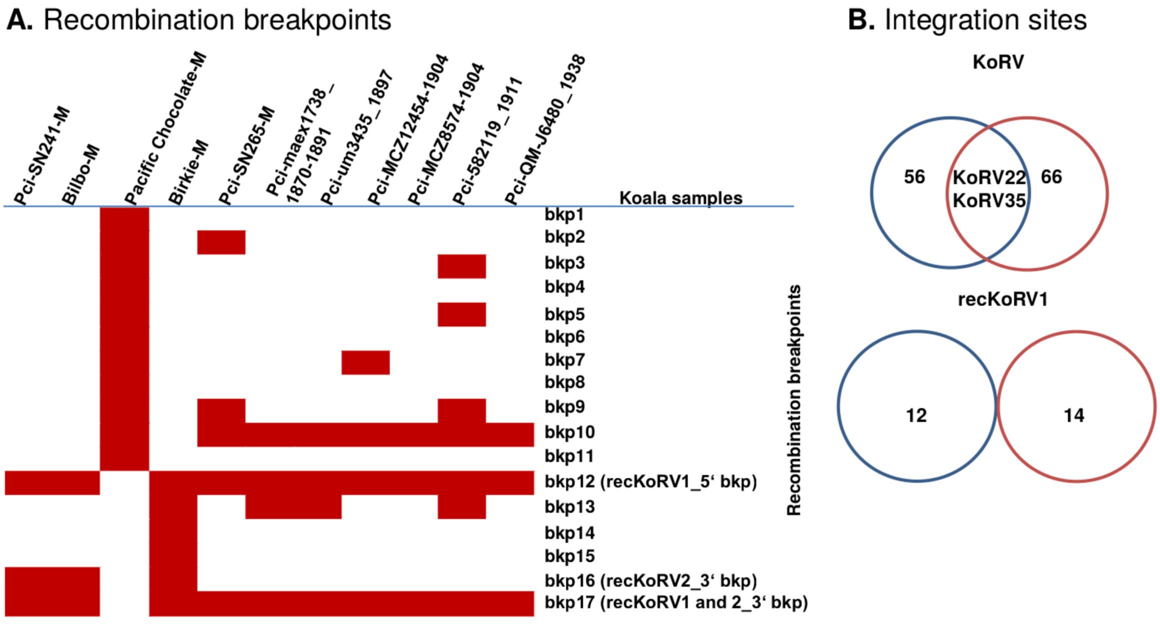


Figure 4.8: Recombination Breakpoints and Integration Sites Across Koalas [2]

Panel A represents as a heatmap presence (red) or absence (white) data from table 4.4, representing identification of recKoRV1 breakpoints (bkp) among koalas: in the iPCR PacBio sequenced data "Bilyarra", the koala reference genome "Bilbo" (KGC), Illumina sequenced koala genomes ("Pacific Chocolate", and "Birke") and hybridization capture data from Tsangaras et al. 2014 [123]. All koalas are from Queensland except for "Pacific Chocolate" who derives from New South Wales. Zoo koalas in the study are primarily derived from Queensland populations. Modern samples are labeled with "M" while the date of collection is indicated for museum samples. Recombination breakpoints (bkp) 1-17 are designated on the vertical axis, with those present in recKoRV1 -3 specially indicated. In panel B, Venn diagrams indicate the degree of overlap of KoRV and recKoRV1 integrations between "Bilbo" (blue, the reference genome from a Queensland koala) and "Bilyarra" (red ; from the Vienna Zoo); only 2 of 120 KoRV integration sites were shared between the two koalas, and none of the 26 recKoRV1 integration sites were shared. The full sequences and reference genome locations for all shared and unique integrations are in table 4.4.

Absence of Reciprocal Recombinant recKoRVs

We screened for reciprocal recombination products relative to the structure of recKoRV1 i.e., containing PhER sequences flanking KoRV coding sequences, and found no evidence for them across the genome of "Bilbo". This was not unexpected because viral integrases are generally LTR sequence-specific and sequence alignment using BLASTn with tolerant/permissive parameter settings revealed no substantial sequence similarity in the LTR regions of KoRV and PhER. Because PhER does not code for an intact integrase, both KoRV and recKoRV1 would rely on KoRV integrase to insert into the genome. PhER-flanked reciprocal recombinants would likely lack the requisite LTR sequences to be recognized and integrated efficiently [203].

Table 4.4: recKoRVs Identified in the Current Study and their Distribution in Modern and Historical Koala Genomes [2]

Breakpoint name*	Individual	Breakpoint orientation†	KoRV orientation	Breakpoint position in KoRV genome**	Breakpoint position in PhER genome***	Microhomology	Recombination breakpoint sequence‡	# reads mapping to bp	Pci-SN265	Pci-QM-J6480	Pci-um3435	Pci-MCZ12454	Pci-MCZ8574	Pci-maex1738	Pci-582119
bp1	Pacific Chocolate	KoRV->PhER	reverse	6019	4519		gccaggtgagagtcatgggttggtgagggttgttatgcAGATGTGTATAAGAGACAGCAATAATAGGACAAAGGTGTAATGTGATAGATGTCTAATTAAAT	1							
bp2	Pacific Chocolate	PhER->KoRV	reverse	504/8430	4628		AGAAAAGCAATAGGATTAGACTGATATGATTGGCTTCCAAGATGTATtgaaagatcccaatgttcgggtagtccaccgaccttgagaaaccctcccaggat	6	X						
bp3	Pacific Chocolate	KoRV->PhER	reverse	3/7929	4625		ggaatgatttctgcctcatgatttctgcctccttcaGTATCTATCATAGCAATCCAAGCTTTAAGCATGTTTAGGTAGGAAGATCACAGTTTTGGATTTT	7							X
bp4	Pacific Chocolate	PhER->KoRV	reverse	504/8430	5580		GGGGGATTGGCAGTATGGCTCCGGCTCTTATGGCAGTCCTCTTCCTTGGATGATTCtgaaagaccccaaagttcgggtagtccaccgaccttgagaaac	4							
bp5	Pacific Chocolate	KoRV->PhER	reverse	3/7929	5577		cgggtagtttccatactccacggaatgatttctgcctcatgatttctgcctccttcaATTCCTTTGCCAAAAACTAAAAGGGGGGAGTAGGAACTCAGGG	7							X
bp6	Pacific Chocolate	KoRV->PhER	forward	504/8430	7279		ggagttgtgttcgcgttgatcctgggagggtttctcaaggtcggtggactacccgaactttgaggtctttcaCCCCATGGTTAACCACTCTTATTTCTGC	3							
bp7	Pacific Chocolate	KoRV->PhER	forward	504/8430	30/7554		gttgatcctgggagggtttctcaaggtcggtggactacccgaactttggggtctttcaATTTTACCAGGTAGTAGAACCCTTTAGGCCATGGAAAACTCCA	4				X			
bp8	Pacific Chocolate	PhER->KoRV	forward	7448	80/7604	CAAGAC	GGTAGTAGAACCCTTTAGGCCATGGAAAACTCCAGATTCCTTAGcaagactccataagtaaactagaggattccttaacctccctgtctgaagtagtgctc	19							
bp9	Pacific Chocolate	KoRV->PhER	forward	3471	106/7630	CCCTC	ctcggctaccgcaaggattcaagaactctcccaccctcAGATCATATCTCTGGAAGCCTCCATTGTGAGAAAGTTTGCAGAGCATTATCATATCTACACAG	1	X						X
bp10	Pacific Chocolate	KoRV->PhER	forward	1395	238/7765	CCCCTCC	ggagacggacagcctgttcctcctcgcagaaccccgccctatccaacatcccctccCTGTCCTGTATACTCCCTATATAAACCCTAGCTCAA	22	X	X	X	X	X	X	X
bp11	Pacific Chocolate	PhER->KoRV	forward	5263	278/7806		AAGATCCATCCCCCTCCCTGTCCTGTATACTCCCTATATAAACCCTAGGACAGAacaaaaacagagacagccctgaccgtctgcaagaagatactagagg	5							
bp12 (recKoRV1, 2 and 3 5' breakpoint)	Birke	KoRV->PhER	forward	1177	3109	AGGAGACT	ttaatccttctattatttctgcagtcaaaaggattgtcttccaggagactgTTGGACCTGGTGAAAGGGTCTAGAAAGAAGAAAGATTTTTTTATTGGGTC	808	X	X	X	X	X	X	X
bp13	Birke	KoRV->PhER	forward	504/830	1/7524		tcctgggagggtttctcaaggtcggaggactacccgaacattggggtctttcaATGTAATGCAGAACATTATTTTACCATTTGGGTAGTAGAACCCTTTAG	2			X			X	X
bp14	Birke	PhER->KoRV	forward	3/7929	74/7598		TAGGTAGTAGAACCCTTTAGGCCATGGAAAACTCCAAATTCCTTAGtgaaggaggcagaaatcatgaggcagaaatcattccgtggagtatggaaactacc	16							
bp15	Birke	KoRV->PhER	forward	506/8431	67/7591		ctgggagggtttctcaaggtcggaggactacccgaacattggggtctttcattAGCAAGACCCTTATCTGGCTGCTCAGGCTCCAGCCCTCAGATCATATC	19							
bp16 (recKoRV3 3' breakpoint)	Birke	PhER->KoRV	forward	36/7962	386/7914		TAATAGCTAATAAAGTCCACAATTTAAAATTAGCTCGGGTCGTCAACTCGCtccgtggagtatggaaactacccggagggcccaaggtttagggacaggtg	60							
bp17 (recKoRV1 and 2 3'breakpoint)	Birke	PhER->KoRV	forward	7619	384/7912	AGGAGACT	TTTAAAATTAGCTTGGGTCGTCAACTCGCCTATAGGGATGAGAGTGGCCCCaggagactcaaggaaaggttagataagaggcagttagagcaccaaaagaa	660	X	X	X	X	X	X	X

* Breakpoints 12, 16 and 17 are found in recKoRV1, 2, or 3

† Breakpoint orientations indicate whether KoRV or PhER ar 5' or 3' of the recombination breakpoint

** 8431 bp; GenBank accession AF1517944. Two positions given when breakpoint is within LTR.

*** 8031 bp; positions 10912078-10920108 of scaffold phaCin_unsw_v4.1.fa.scaf00062 (NCBI reference sequence NW_018344013.1) from phaCin_unsw_v4.1 genome assembly. Two positions given when breakpoint is within LTR.

‡ KoRV sequences are shown in black and PhER sequences in red capitalized letters

Koalas in dark grey are modern and those in light grey are historical

4.2.4 Comparison of LTRs and Integration Sites Among Koala Genomes

KoRV integration sites are highly insertionally polymorphic across unrelated koalas [119, 141]. We examined KoRV and recKoRV1 integration sites in “Bilyarra” using a long read iPCR strategy and PacBio sequencing. This method allowed for identifying KoRV and recKoRV1 sequences and their integration sites in long single PacBio reads. “Bilyarra” exhibited a higher number of KoRV integration sites (66) than found in the “Bilbo” reference genome (58). Among the KoRV integration sites (unique to each proviral locus) in “Bilbo” and “Bilyarra”, only two were shared (KoRV22 and KoRV35; figure 4.8 (B) and tables A.1 A.2, and A.3). In each of the two KoRV proviral loci shared by “Bilbo” and “Bilyarra”, deletions detected in the envelope gene (*env*) would likely have precluded production of infectious virions. The other 120 integration sites were only detected in one of the two koalas. This suggests that individual KoRV integrants are found at low frequencies in their respective chromosomes and not generally shared by unrelated koalas. By contrast, many LTR sequences from “Bilbo”, “Bilyarra”, and other koalas were identical; they largely overlapped across a minimum spanning network, with few sequences unique to a specific koala (4.9). This indicates that KoRV proviruses at different loci have the same LTR sequence [141], as many LTR sequences, unlike integration sites, were shared among unrelated koalas.

Twenty-four recKoRV integrations were identified in “Bilyarra”, of which 14 were characterized as recKoRV1. An overview of shared target site duplications in “Bilyarra” is given in table 4.5. In “Bilbo” 12 recKoRV1 sites were identified (identified through PacBio sequence reads that included both the integration sites and one or both of the recKoRV1 breakpoints). None of the recKoRV1 integration sites was shared between “Bilbo” and “Bilyarra” (figure 4.8 (B)). This may indicate that recKoRV1 integrations have not had sufficient time to become broadly distributed among koalas. The absence of shared loci carrying recKoRV1 between “Bilbo” and “Bilyarra” would suggest that recKoRV1 has been able to retrotranspose to different loci in the koala genome and/or that the same recombination event has occurred between KoRV and PhER on more than one occasion. The recKoRV LTRs varied across recKoRV1 loci, were often identical to KoRV LTRs, and included four of the five most common KoRV LTRs (4.9). This suggests that the same breakpoints between KoRV and PhER have been used in independent recombination events to generate recKoRV1s multiple times, because random mutations from a single ancestral recKoRV1 LTR would not exactly match those that happen to distinguish the most common KoRV LTR sequences. Target site duplications of 4-10 bp were detected at the integration sites of KoRV and recKoRV1 in the large majority of cases (as shown in supplement A.1 tables A.1, A.2 and A.3), suggesting that the integrations involved a retrovirus-typical reverse transcription as opposed to meiotic recombination.

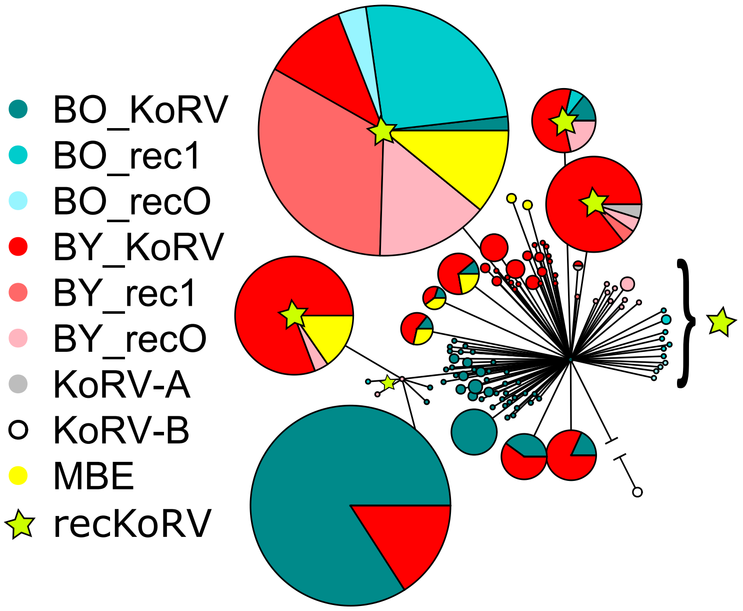


Figure 4.9: Relationships Among KoRV LTR Sequences [2]

A minimum spanning network shows the relationships among the KoRV, recKoRV1 (rec1) and other recKoRV (recO) LTR sequences identified in koalas “Bilyarra” (BY), “Bilbo” (BO) and koalas examined in reference [141] (MBE). Each pie chart represents a distinct LTR sequence, with circle sizes proportional to the frequency of occurrence of each sequence. Alternative relationships among the network are shown as grey lines. The poor resolution among LTR groups is due to the low diversity among individual elements often differing by single nucleotide differences.

Table 4.5: Shared Target Site Duplications “Bilyarra”

Location	Type	Orientation	Position	Sequence	Reverse Complement**
scaf00234	recKoRV1	forward	4,233,389	AAGAT	
scaf00139	recKoRV*	forward	4,264,851	AAGAT	
scaf00043	KoRV	forward	10,816,237	AGAT	
scaf00053	KoRV	reverse	5,206,411	ATCT	AGAT
scaf00275	recKoRV1	forward	2,910,435	AGGT	
scaf00014	recKoRV1	reverse	15,030,324	ACCT	AGGT
scaf00035	recKoRV*	forward	5,731,424	ATGG	
scaf00002	KoRV	reverse	25,542,095	CCAT	ATGG
scaf00441	KoRV	forward	282,050	ATTC	
scaf00048	KoRV	reverse	8,799,477	GAAT	ATTC
scaf00218	KoRV	ND	1,849,618	CAAT	ATTG
scaf00013	KoRV	reverse	14,937,508	TAGG/CAAT	CCTA/ATTG
scaf00094	KoRV	reverse	1,633,494	GAAA	TTTC
scaf00098	KoRV	reverse	2,205,920	GAAA	TTTC
scaf00354	KoRV	forward	1,495,511	GAGC	
scaf00164	KoRV	reverse	5,577,461	GCTC	GAGC
scaf00634	KoRV	reverse	18,365	GCTC	GAGC

* Undetermined recombinant KoRV (not recKoRV1)

** The reverse complement is only computed for viral integrants with reverse orientation

4.2.5 Recombination Breakpoint Distributions Among Koala Populations

Unlike most ERVs, KoRV greatly varies in prevalence across its host populations. While all Queensland koalas are positive for KoRV with high copy numbers in their genomes, southern Australian koalas have a much lower prevalence and copy number, with KoRV completely absent from some individuals [204]. Sequences for all 17 recombination breakpoints identified between KoRV and PhER were used to query the koala reference genome and Illumina sequence datasets. Of the 17, eleven were identified in the genome of ‘Pacific Chocolate’, a koala from New South Wales, but were absent from the genome of “Birke” (table 4.1) from Queensland. The other six recombination breakpoints, including the breakpoints of recKoRV1, were identified in “Birke” but absent from ‘Pacific Chocolate’. The lack of overlap may suggest that independent recombination events between KoRV and PhER have occurred in koalas from the two Australian regions. Screening of sequence datasets that had been generated after hybridization capture of KoRV identified the two recKoRV1 recombination breakpoints in all other koalas examined, including both museum and modern samples (figure 4.8 and table 4.4). Five of the 11 breakpoints in ‘Pacific Chocolate’ from New South Wales were specific to that individual. The remaining six breakpoints were detected only sporadically among existing Illumina datasets, with the exception of breakpoint 10, which was found in most museum and a zoo koala but not in “Birke”.

4.2.6 An Extended Analysis of the Geographic Distribution of recKoRV1

To more precisely characterize the geographic distribution of recKoRV1 among koalas, the presence or absence of the 3' recKoRV1 recombination breakpoint was examined using PCR. To span the recombination breakpoint, the 5' PCR primer matched the upstream PhER sequence and the 3' primer matched the downstream KoRV LTR sequence. We screened for the 3' recKoRV1 recombination breakpoint in 166 koalas from 11 populations across Australia that had previously been screened for KoRV prevalence (17). KoRV and the recKoRV1 3' breakpoint were both present across all koalas in Queensland and inland New South Wales (figure 4.10) with the notable exception of St. Bees Island (figure 4.10, population B) in which the recKoRV1 3' breakpoint was only detected in 4 of 15 koalas, although KoRV was ubiquitous among St. Bees koalas. The coastal population of Port Stephens in New South Wales (figure 4.10 population G) was 100 percent KoRV positive but devoid of recKoRV1. This is consistent with the absence of recKoRV1 recombination breakpoints in the genome of 'Pacific Chocolate' (from nearby Port Macquarie, New South Wales). Further south in Victoria, both Mornington Peninsula and Gippsland koalas were negative for the recKoRV1 breakpoint and either positive for KoRV (Gippsland; figure 4.10, population K) or negative for both KoRV and recKoRV1 (Mornington Peninsula; figure 4.10, population J).

4.3 Discussion

Degradation of ERV genomes, and the loss of *env* in particular, may benefit the host by preventing the production of virulent retroviruses that can spread horizontally [192]. Our findings suggest that the recombination-mediated degradation of retroviruses, which has been postulated for many human and other vertebrate ERVs, and the genomic proliferation of recombinants both occur at the earliest stages of retroviral germ-line invasion [205, 206, 207, 208]. This is supported by the presence of recKoRV1s in koalas across almost all of Australia in both modern and historical samples, and their high copy number in the koala genomes examined (figures 4.2 and 4.10). KoRV is thought to have invaded the koala germline relatively recently, within the last 50,000 y [141]. Thus, within this time frame, the recKoRVs were generated and recKoRV1s arrived at the widespread distribution revealed here.

Seventeen recombination breakpoints were detected between KoRV and PhER. Recombination occurred in some cases at microhomologies, short sequences common to the two retroelements that likely enabled recombination at many of the breakpoints, including those of recKoRV1. Transcripts of PhER have been detected in the koala transcriptome, suggesting that PhER could be copackaged with KoRV in the same virion [122], enabling recombination. KoRV integrants may also have recombined with retrotranscribed PhER during meiosis. In both KoRV and recKoRV1, target site duplications were generally detected in the host genome flanking the 5' and 3' ends of the provirus, indicating that recombination between retroelements at different loci had not affected these integrants. Such recombinants, which can delete large regions of the genome, may have been removed by selection. Once a recKoRV is established in the germline, it can spread vertically (and geographically) across koala populations.

A high degree of population structuring was detected among the different recombination breakpoints between KoRV and PhER. In particular, the recKoRV1 3' breakpoint

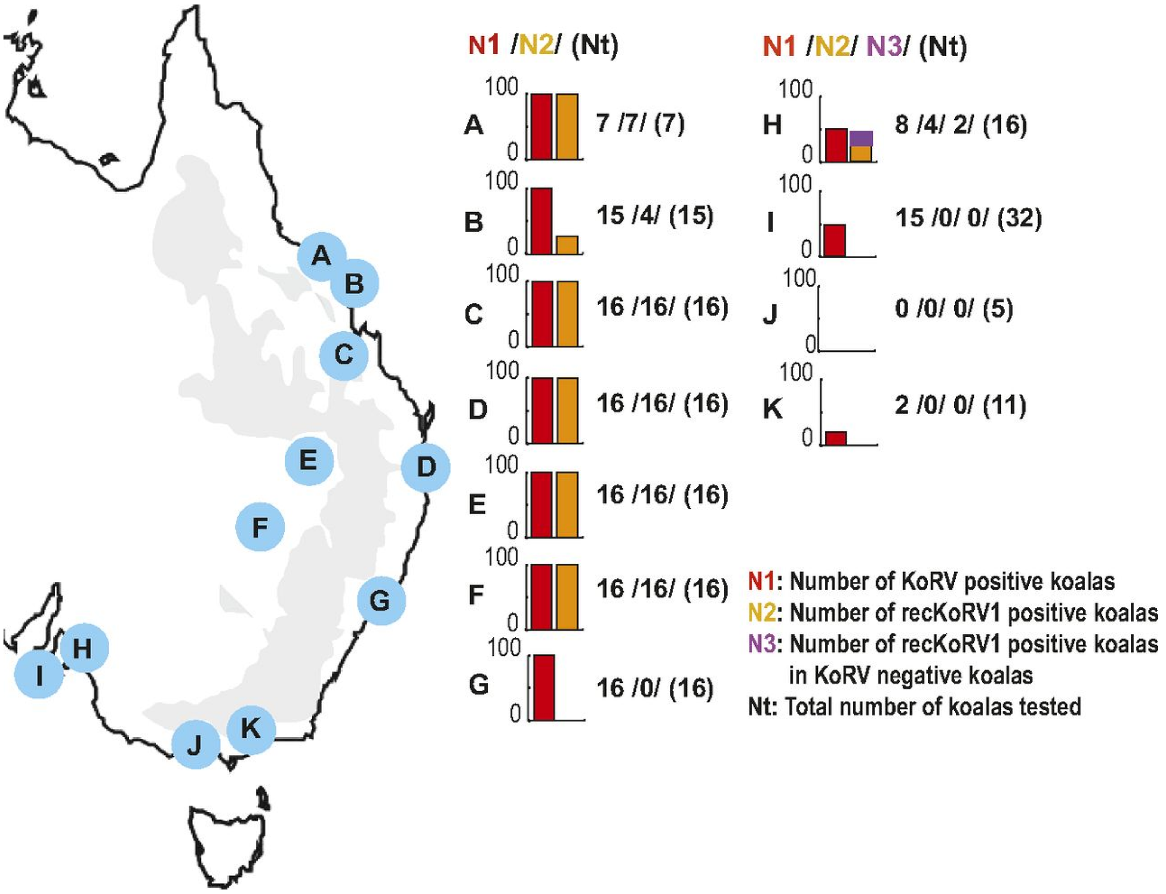


Figure 4.10: Prevalence of recKoRV1 in KoRV-positive and KoRV-negative Koalas Across Australia

The proportion of recKoRV1-positive koalas in both KoRV-positive and KoRV-negative koalas was determined by PCR assay. The percent of KoRV-positive and KoRV-negative koalas with or without recKoRV1 is shown for each population in the bar charts. The numbers to the right of each chart indicate the number of koalas in each respective category (N1, N2, and N3). Nt (in parentheses) refers to the total number of koalas tested at each locality. Red bars on the graphs indicate the percent of koalas that were KoRV positive, orange indicates the percent recKoRV1 positive, and purple indicates the percent of koalas recKoRV1 positive but KoRV negative. The Great Dividing Range is indicated on the map in gray. The localities sampled were as follows: A, Hamilton Island, Queensland (QLD); B, St Bees Island, QLD; C, Central QLD; D, Currumbin Wildlife Sanctuary, QLD; E, South-West QLD, F, West Pilliga, New South Wales (NSW); G, Port Stephens, NSW; H, Adelaide Hills, South Australia (SA); I, Kangaroo Island, SA; J, Mornington Peninsula, Victoria (VIC); K, Gippsland, VIC.

was completely absent from some populations in New South Wales (figure 4.10), while the genomes of two koalas, one from Queensland and one from coastal New South Wales, differed dramatically in their complement of recombination breakpoints (figure 4.8). Genetic differentiation between Queensland and New South Wales koalas has been reported in previous studies [209, 210], suggesting restricted gene flow between koalas in the two states, perhaps in part due to the Great Dividing Range (figure 4.10). The barrier to gene flow cannot be complete because KoRV is present at high frequency in all of New South Wales and was thus likely transferred from koalas in Queensland at some point. Additionally, koala populations do not show high degrees of genetic structure compared with other marsupials, although recent barriers to gene flow may exist particularly in New South Wales [211, 212]. We cannot rule out the possibility that regional differences in PhER expression may affect the genesis or distribution of recKoRVs by altering the amount or type of PhER template available for recombination. However, it is also possible that PhER and KoRV may be expressed and recombine in any population where both are present. This is supported by the analysis of KoRV-PhER recombination breakpoints in the genomes of a koala from Queensland and a koala from New South Wales. The two carried completely distinct sets of recombination breakpoints (figure 4.8), suggesting that recombinants between KoRV and PhER formed independently in the two populations.

Several populations showed atypical patterns in the distribution of recKoRV1. In St. Bees Island off Queensland, only 4 of 15 koalas were recKoRV1 positive, but all were KoRV positive (figure 4.10). This contrasts with mainland Queensland for which all koalas tested (n=48) were positive for both recKoRV1 and KoRV. The St. Bees Island population was founded by translocation of 12–17 koalas from mainland Queensland in the 1930s [213]. The founding population of St. Bees was small, likely with insertional polymorphisms in each recKoRV1 locus. After the population expanded, the koalas would reflect random combinations of the small numbers of founder chromosomes. It may be that loci carrying recKoRV1 were randomly lost through genetic drift, although it is also possible that selection may have played a role.

In the Adelaide Hills of South Australia, several KoRV-negative individuals proved to be recKoRV1 positive (figure 4.10). KoRV copy number has been shown to decrease dramatically in southern Australia based on qPCR targeting the *pol* gene, and KoRVs rarely exist in both chromosomes in a given koala individual even where copy numbers are high [34, 141]. The recKoRV1-positive individuals lacking KoRV likely reflect Mendelian segregation of integrants in a population where both KoRV and recKoRV1s are present at low copy numbers and at low frequencies at their respective loci, so that only a limited proportion of individuals carry either or both.

KoRV would suffer the loss of virulence after recombination with PhER because none of the recombinants are predicted to code for an intact virus. Existing genomic elements like PhER would proliferate by having parts of their sequences incorporated into recKoRVs. While recKoRVs could still potentially exert deleterious effects on the host, e.g. by retrotransposition into new genomic locations, other potentially deleterious effects of the provirus would be reduced relative to intact KoRV, notably the ability of these elements to produce infectious retrovirus. The switch to a proviral form that is disrupted by recombination may be one aspect of the transition from horizontal to vertical transmission among ERVs. Over time, this would be expected to result in an increase in recKoRV abundance at the expense of virulent KoRV proviruses, potentially reducing the impact of the latter. The pressure to make this transition may be higher in long-lived species that are more likely to be affected by ERVs with oncogenic potential [214]. During the transition period when infectious KoRV and recombinants coexist, KoRV particles may de novo generate and horizontally transmit recKoRVs (figure 4.9) and in this manner coinfect host cells, although superinfection resistance would likely limit the novel infection-mediated proliferation of both KoRV and recKoRVs [12].

In detecting large numbers of recombinants between KoRV and PhER, we establish that recombination with existing retroelements may be one way in which the ability of retroviruses invading the germline to faithfully replicate is disrupted, by removing their ability to encode active viruses associated with disease. This would not be the only mechanism by which a host species controls an invading ERV, since other factors are likely to play a role, such as methylation or antiretroviral proteins or disruptive within-KoRV recombination (as was evident for KoRV22 and KoRV35, the only shared KoRV integrants identified) [215]. Nor would recKoRV lack potentially deleterious aspects of a provirus, as activation or disruption of genes at or near insertion sites may still occur. However, the deleterious effects of recKoRVs are not likely to be as great as those of KoRVs, and recKoRVs may thus be less subject to purifying selection than replication competent KoRVs, allowing recKoRVs to persist in the host germline. Several lines of evidence suggest that production of recKoRVs may reflect a general means of accommodation between ERV and host. The recKoRV proviruses would have a reduced ability to proliferate relative to intact KoRV. The process of recKoRV formation has occurred frequently and independently, given the many recKoRVs identified and geographic differences in the occurrence of breakpoints. The degraded nature of recKoRV1 is also consistent with inferences drawn from more ancient ERVs in vertebrate genomes, notably the concept of genomic superspreaders, which suggests that retroviruses that lose the *env* gene will be more successful at propagating in host genomes than intact ERVs [192]. It is also consistent with the exchange of sequences between divergent retroviruses, which has been inferred for ERVs in various host species ([205, 206, 207, 208, 216]). For example, some human ERVs are believed to have recombined before their proliferation [208]. This suggests that recombination-based degradation has occurred during invasions of vertebrate germlines by different groups of retroviruses. Our study demonstrates empirically that the generation of such recombinants occurs during the early stages of genomic invasion by ERVs of a host germline.

4.3.1 recKoRV Classification

In figure 4.11 a mapping of "Bilyarra" sequences determined as recKoRV integration site at position 10,330,590 of scaffold 3 in the koala genome is shown (list of all recKoRV integration sites see table A.3). Even we had clear evidence, at least from the 3' integration site, that this is a recKoRV1 locus, we were not able to amplify and Sanger sequence this locus. This clearly shows that the number of recKoRV1s might be underestimated in this study. Additional information about results from the amplification of candidate loci and Sanger sequencing are shown in the appendix, figures A.1 to A.11 and tables A.5 and A.4.

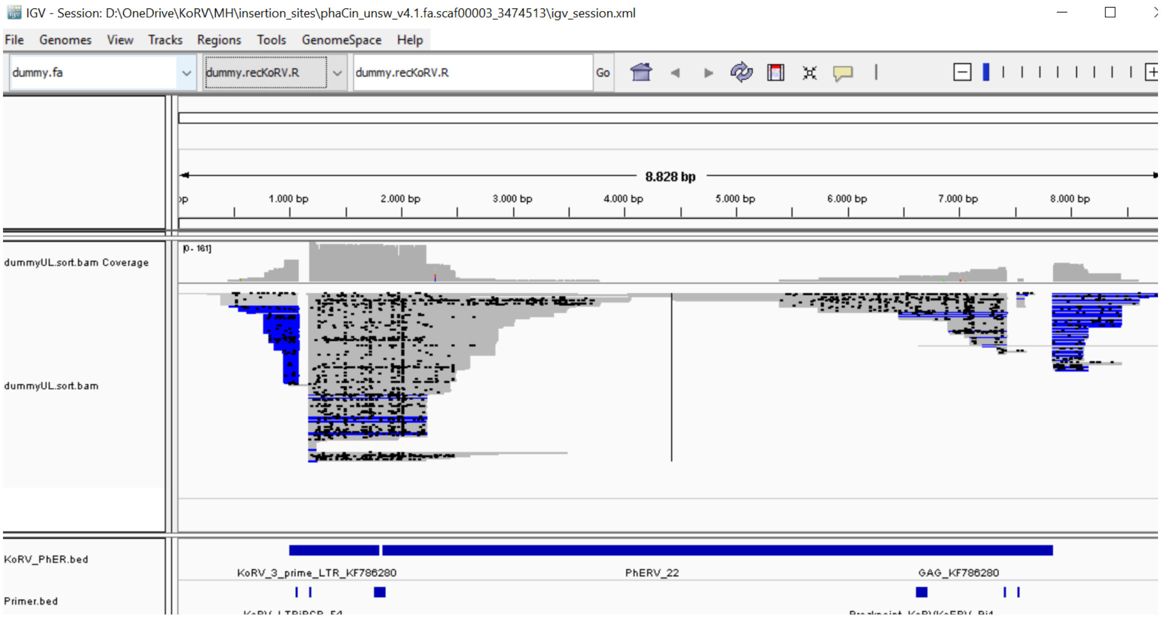


Figure 4.11: Mapping Result for Scaffold 3 recKoRV Reverse

Example of recKoRV1 candidate which did not amplify trying to validate using Sanger sequencing. (describe figure)

Chapter 5

Conclusions

The development of algorithms, tools, and pipeline to process large data sets ("big data") from long-read sequencing technologies is an important issue for future research. The prospect of establishing easy-to-use software, enabling users without a computational background to analyze the huge amounts of data which are produced every day is of the fundamental need to exploit the potential of sequencing technologies.

My main objective was to investigate methods to process data from targeted short and long read high-throughput sequencing (HTS) to comprehensively compare integration sites of endogenous retroviruses.

5.1 Resume

The evidence from chapter 2 suggests that a combination of Primer Extension Capture (PEC) and Hybridization Capture (HC) reveals the highest target enrichment efficiency for retroviral profiling from ancient DNA. Even though there are still limitations regarding a 3' and 5' integration site bias for every method tested, we were able to establish methods achieving the objective. An automated computational pipeline was developed to profile koala retroviruses in historical samples comprehensively. The pipeline is suitable for any target enrichment experiment using PEC, Single Primer Extension (SPEX) or HC combined with short-read sequencing. The most critical limitation is due to the memory intensive and time-consuming database search and clustering. While the efficiency of alignment methods has improved, further tests are needed to automatically estimate the best granularity of clusters, which has been evaluated manually for the study.

In chapter 3 I described a new method to investigate retroviral integration sites making use of long-read sequencing technologies. As displayed, we have succeeded in retrieving koala retrovirus (KoRV) integration sites up to 6 kb in length. We established an innovative approach for restriction-enzyme-free target enrichment of endogenous retroviruses. Using SiP, we produced comprehensive results providing new insights into the evolution of the koala retrovirus discussed in chapter 4.

The findings of chapter 2 and chapter 4 are in accordance and support that the prediction that the proportion of integration sites shared between any two koalas is relatively small. The performance of adapter ligation in SiP was rather disappointing. This benchmarks the potential loss of sequencing products during standard library preparations for short-read sequencing, however, it would have been a feasible tag for post-processing and restructuring inverse PCR reads. Surprisingly, every protocol for capturing retroviral integration sites was biased towards either the 3' or 5' integration site. It remains unclear, why these biases occurred.

We found and described a novel recombinant, designated recKoRV1 and 15 other recombinants of an older retroelement and KoRV. These findings support the hypothesis that older degraded retroelements might tame replication-competent endogenous retroviruses (ERVs) during the earliest stages of germline invasion. Long-read sequencing technologies offer new opportunities to profile retroviral integration sites. However, application areas for short-read sequencing technologies still exist. Highly fragmented DNA such as ancient DNA will not be suitable for long-read sequencing. Thus short-read sequencing will be the method of choice. The question of whether long-read sequencing will be available at lower costs than short-read sequencing remains open but will change the range of applications as well. Since circular consensus sequences reach consensus accuracies of 99.9 % [217], sequencing error rates similar to short-read sequencing error rates might be achieved in the future. The obvious benefits of long-read sequencing are sequencing through repetitive elements, variant phasing, detection of methylation patterns and a uniform coverage [218]. Moreover, a combination of short- and long-read sequencing can be used to combine the advantages of both techniques. Hybrid approaches correcting low-quality long reads with short reads have been developed and recently used for the Vertebrate Genome Project [219].

5.2 Perspectives/Outlook

I propose that future research should be undertaken in the following areas:

1. Establishment of a database for retroviral integration sites
2. Development of tools which automatically restructure inverse PCR reads
3. Assess adapter ligation efficiency
4. Development of software packages for classification of endogenous retroviruses
5. Examine a molecular clock to approximate endogenization events

To the best of my knowledge, there is no common database for retroviral integration sites, except the Retrovirus Integration Database (RID) [220], which is limited to HIV1, HTLV1, MLV, ALV/Hg19, mm9, Gal4. Inverse PCR is a very promising method for the retrieval of retroviral integration sites. Even though we managed to process the data and obtain our objectives, SiP might help to suggest several courses of action in order to profile endogenous retroviruses in non-model organisms. Unfortunately, I eventually had to use a host reference genome in my pipeline. We tested adapter ligation efficiency under various conditions using only one sample. This research has raised many questions in need of further investigation. First, is the loss of sequencing products due to inefficient adapter ligation biased? Second, is it possible to increase the adapter ligation efficiency to achieve more reasonable results? Third, can we use adapters as a flag for automatic processing inverse PCR products?

All pipelines developed are proof of concept. It would be of fundamental interest to create packages which implement the designed pipelines for reusability. This would enable researchers to either process short read HTS products as well as long read HTS products from target enrichment protocols for ERV integration sites. Initially, we presumed that dating endogenization events would be possible by the comparison of target site duplications from insertion sites. In contradiction with earlier findings [141], we found that the length of target site duplications was not limited to four base pairs, but varied from four to ten base pairs. Due to this fact, extensive investigations are needed in order to develop methods for reliable dating of endogenization events.

Returning to the hypothesis posed at the beginning of this thesis, further studies are needed to investigate KoRV integration sites in tumor and control tissues to assess the link between endogenous retroviruses and the disease status. Further studies, which focus on the comparison of KoRV integration sites in relatives, to directly investigate the inheritance of endogenous retroviruses are undertaken.

Appendix A

Appendix

The 166 wild koala DNA samples used to define the distribution of the recKoRV1 3' breakpoint across Australian koala populations were collected by Joanne Meers, Paul Young and their associates [34]. Screening of these samples was performed on DNA extracted using the Blood & Tissue DNA Extraction Kit (Qiagen) or from DNA provided by collaborators and tested for integrity using a control primer pair-specific for the koala actin gene. The study was conducted in accordance with the following permits and approvals: the University of Queensland Animal Ethics Committee (approval numbers SVS/492/12/ARC/WWW; SVS/488/09/ARC/WWW and MICRO/-PARA/612/08/ARC); the South Australia Department of Environment, Water and Natural Resources Permit to Undertake Scientific Research (permit numbers A25 844; U25790); the South Australia Department of Environment, Water and Natural Resources Wildlife Ethics Committee (approval numbers 12/2010; 51-2009-M1); the South Australia Department of Environment, Water and Natural Resources Export Protected Animals Permit (permit number E20833); Queensland Environmental Protection Agency Wildlife Movement permits (numbers WIWM08219010; WIWM09103211; WIWM09434211; WIWM12645213; WIWM06555009; WIWM06798010), NSW National Parks and Wildlife Service Export Licence, number IE106347.

Bibliography

[1] P. Cui, U. Löber, D. E. Alquezar-Planas, Y. Ishida, A. Courtiol, P. Timms, R. N. Johnson, D. Lenz, K. M. Helgen, A. L. Roca, S. Hartman, and A. D. Greenwood, "Comprehensive profiling of retroviral integration sites using target enrichment methods from historical koala samples without an assembled reference genome," *PeerJ*, vol. 4, p. e1847, 2016.

[2] U. Löber, M. Hobbs, A. Dayaram, K. Tsangaras, K. Jones, D. E. Alquezar-Planas, Y. Ishida, J. Meers, J. Mayer, C. Quedenau, W. Chen, R. N. Johnson, P. Timms, P. R. Young, A. L. Roca, and A. D. Greenwood, "Degradation and remobilization of endogenous retroviruses by recombination during the earliest stages of a germ-line invasion," *Proceedings of the National Academy of Sciences*, p. 201807598, Aug. 2018.

[3] D. E. Alquezar-Planas, U. Löber, P. Cui, C. Quedenau, W. Chen, and A. D. Greenwood, "SIP: DNA Sonication Inverse PCR for Genome-Scale Analysis Integration Sites," *TBA*, Mar. 2019.

[4] A. Hayward, "Origin of the retroviruses: when, where, and how?," *Current Opinion in Virology*, vol. 25, pp. 23–27, Aug. 2017.

[5] T. H. Bestor, "The host defence function of genomic methylation patterns," *Novartis Foundation Symposium*, vol. 214, pp. 187–195; discussion 195–199, 228–232, 1998.

[6] I. C. on Taxonomy of Viruses (ICTV), "Virus Taxonomy: 2017 Release," Mar. 2018.

[7] J. M. Coffin, S. H. Hughes, and H. Varmus, *Retroviruses.* Plainview, N.Y.: Cold Spring Harbor Laboratory Press, 1997.

[8] J. Demeulemeester, J. De Rijck, R. Gijsbers, and Z. Debyser, "Retroviral integration: Site matters," *Bioessays*, vol. 37, pp. 1202–1214, Nov. 2015.

[9] S. Desfarges and A. Ciuffi, "Retroviral Integration Site Selection," *Viruses*, vol. 2, pp. 111–130, Jan. 2010.

[10] C. Petropoulos, *Retroviral Taxonomy, Protein Structures, Sequences, and Genetic Maps.* Cold Spring Harbor Laboratory Press, 1997.

[11] G. N. Maertens, S. Hare, and P. Cherepanov, "The mechanism of retroviral integration from X-ray structures of its key intermediates," *Nature*, vol. 468, pp. 326–329, Nov. 2010.

[12] M. Nethe, B. Berkhout, and A. C. van der Kuyl, "Retroviral superinfection resistance," *Retrovirology*, vol. 2, p. 52, 2005.

[13] S. P. Goff, "Retrovirus Restriction Factors," *Molecular Cell*, vol. 16, pp. 849–859, Dec. 2004.

[14] P. Klenerman, H. Hengartner, and R. M. Zinkernagel, "A non-retroviral RNA virus persists in DNA form," *Nature*, vol. 390, pp. 298–301, Nov. 1997.

[15] R. Castanera, L. López-Varas, A. Borgognone, K. LaButti, A. Lapidus, J. Schmutz, J. Grimwood, G. Pérez, A. G. Pisabarro, I. V. Grigoriev, J. E. Stajich, and L. Ramírez, "Transposable Elements versus the Fungal Genome: Impact on Whole-Genome Architecture and Transcriptional Profiles," *PLoS Genetics*, vol. 12, June 2016.

[16] J. H. Grau, A. J. Poustka, M. Meixner, and J. Plötner, "LTR retroelements are intrinsic components of transcriptional networks in frogs," *BMC Genomics*, vol. 15, p. 626, July 2014.

[17] A. Lee, A. Nolan, J. Watson, and M. Tristem, "Identification of an ancient endogenous retrovirus, predating the divergence of the placental mammals," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 368, Sept. 2013.

[18] J. Denner, "Expression and function of endogenous retroviruses in the placenta," *APMIS: acta pathologica, microbiologica, et immunologica Scandinavica*, vol. 124, pp. 31–43, Feb. 2016.

[19] A. A. Buzdin, V. Prassolov, and A. V. Garazha, "Friends-Enemies: Endogenous Retroviruses Are Major Transcriptional Regulators of Human DNA," *Frontiers in Chemistry*, vol. 5, 2017.

[20] D. Nikitin, D. Penzar, A. Garazha, M. Sorokin, V. Tkachev, N. Borisov, A. Poltorak, V. Prassolov, and A. A. Buzdin, "Profiling of Human Molecular Pathways Affected by Retrotransposons at the Level of Regulation by Transcription Factor Proteins," *Frontiers in Immunology*, vol. 9, Jan. 2018.

[21] J. Kawasaki and K. Nishigaki, "Tracking the Continuous Evolutionary Processes of an Endogenous Retrovirus of the Domestic Cat: ERV-DC," *Viruses*, vol. 10, Apr. 2018.

[22] IUCN, "Phascolarctos cinereus: Woinarski, J. & Burbidge, A.A.: The IUCN Red List of Threatened Species 2016: e.T16892a21960344," tech. rep., International Union for Conservation of Nature, May 2014. type: dataset.

[23] R. M. Nowak, *Walker's Marsupials of the World.* JHU Press, Sept. 2005. Google-Books-ID: ldXtY8ppxSQC.

[24] J. A. W. Kirsch and J. H. Calaby, "The species of living marsupials: an annotated list," in *The Biology of Marsupials*, Studies in Biology, Economy and Society, pp. 9–26, Palgrave, London, 1977.

[25] W. P. Heuschele and J. R. Hayes, "Acute leukemia in a New South Wales koala (Phascolarctos c. cinereus)," *Cancer Research*, vol. 21, pp. 1394–1395, Nov. 1961.

[26] T. Backhouse and A. Bolliger, “Morbidity and Mortality in the Koala (Phascolarctos cinereus).,” *Australian Journal of Zoology*, vol. 9, no. 1, p. 24, 1961.

[27] R. Tarlinton, J. Meers, J. Hanger, and P. Young, “Real-time reverse transcriptase PCR for the endogenous koala retrovirus reveals an association between plasma viral load and neoplastic disease in koalas,” *The Journal of General Virology*, vol. 86, pp. 783–787, Mar. 2005.

[28] C. A. Waugh, J. Hanger, J. Loader, A. King, M. Hobbs, R. Johnson, and P. Timms, “Infection with koala retrovirus subgroup B (KoRV-B), but not KoRV-A, is associated with chlamydial disease in free-ranging koalas (Phascolarctos cinereus),” *Scientific Reports*, vol. 7, Mar. 2017.

[29] U. Fiebig, M. G. Hartmann, N. Bannert, R. Kurth, and J. Denner, “Transspecies Transmission of the Endogenous Koala Retrovirus,” *Journal of Virology*, vol. 80, pp. 5651–5654, June 2006.

[30] J. Denner and P. R. Young, “Koala retroviruses: characterization and impact on the life of koalas,” *Retrovirology*, vol. 10, p. 108, Oct. 2013.

[31] Y. Ishida, K. Zhao, A. D. Greenwood, and A. L. Roca, “Proliferation of endogenous retroviruses in the early stages of a host germ line invasion,” *Molecular Biology and Evolution*, vol. 32, pp. 109–120, Jan. 2015.

[32] K. J. Chappell, J. C. Brealey, A. A. Amarilla, D. Watterson, L. Hulse, C. Palmieri, S. D. Johnston, E. C. Holmes, J. Meers, and P. R. Young, “Phylogenetic diversity of Koala Retrovirus within a Wild Koala Population,” *Journal of Virology*, pp. JVI.01820–16, Nov. 2016.

[33] M. Hobbs, A. King, R. Salinas, Z. Chen, K. Tsangaras, A. D. Greenwood, R. N. Johnson, K. Belov, M. R. Wilkins, and P. Timms, “Long-read genome sequence assembly provides insight into ongoing retroviral invasion of the koala germline,” *Scientific Reports*, vol. 7, Nov. 2017.

[34] G. S. Simmons, P. R. Young, J. J. Hanger, K. Jones, D. Clarke, J. J. McKee, and J. Meers, “Prevalence of koala retrovirus in geographically diverse populations in Australia,” *Australian Veterinary Journal*, vol. 90, pp. 404–409, Oct. 2012.

[35] R. A. Weiss, “On the concept and elucidation of endogenous retroviruses,” *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 368, p. 20120494, Sept. 2013.

[36] O. T. Avery, C. M. MacLeod, and M. McCarty, “Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii,” *Journal of Experimental Medicine*, vol. 79, pp. 137–158, Feb. 1944.

[37] R. W. Holley, J. Apgar, G. A. Everett, J. T. Madison, M. Marquisee, S. H. Merrill, J. R. Penswick, and A. Zamir, “STRUCTURE OF A RIBONUCLEIC ACID,” *Science (New York, N.Y.)*, vol. 147, pp. 1462–1465, Mar. 1965.

[38] F. Sanger, S. Nicklen, and A. R. Coulson, “DNA sequencing with chain-terminating inhibitors,” *Proceedings of the National Academy of Sciences*, vol. 74, pp. 5463–5467, Dec. 1977.

[39] A. M. Maxam and W. Gilbert, “A new method for sequencing DNA,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, pp. 560–564, Feb. 1977.

[40] J. M. Heather and B. Chain, “The sequence of sequencers: The history of sequencing DNA,” *Genomics*, vol. 107, pp. 1–8, Jan. 2016.

[41] J. Tost and I. G. Gut, “DNA analysis by mass spectrometry-past, present and future,” *Journal of mass spectrometry: JMS*, vol. 41, pp. 981–995, Aug. 2006.

[42] S. Drmanac, D. Kita, I. Labat, B. Hauser, C. Schmidt, J. D. Burczak, and R. Drmanac, “Accurate sequencing by hybridization for DNA diagnostics and individual genomics,” *Nature Biotechnology*, vol. 16, pp. 54–58, Jan. 1998.

[43] J. Shendure, G. J. Porreca, N. B. Reppas, X. Lin, J. P. McCutcheon, A. M. Rosenbaum, M. D. Wang, K. Zhang, R. D. Mitra, and G. M. Church, “Accurate multiplex polony sequencing of an evolved bacterial genome,” *Science (New York, N.Y.)*, vol. 309, pp. 1728–1732, Sept. 2005.

[44] J. Ju, D. H. Kim, L. Bi, Q. Meng, X. Bai, Z. Li, X. Li, M. S. Marma, S. Shi, J. Wu, J. R. Edwards, A. Romu, and N. J. Turro, “Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, pp. 19635–19640, Dec. 2006.

[45] M. L. Metzker, “Sequencing technologies - the next generation,” *Nature Reviews. Genetics*, vol. 11, pp. 31–46, Jan. 2010.

[46] N. J. Loman, R. V. Misra, T. J. Dallman, C. Constantinidou, S. E. Gharbia, J. Wain, and M. J. Pallen, “Performance comparison of benchtop high-throughput sequencing platforms,” *Nature Biotechnology*, vol. 30, no. 5, pp. 434–439, 2012.

[47] E. L. v. Dijk, H. Auger, Y. Jaszczyszyn, and C. Thermes, “Ten years of next-generation sequencing technology,” *Trends in Genetics*, vol. 30, pp. 418–426, Sept. 2014.

[48] I. Braslavsky, B. Hebert, E. Kartalov, and S. R. Quake, “Sequence information can be obtained from single DNA molecules,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, pp. 3960–3964, Apr. 2003.

[49] E. E. Schadt, S. Turner, and A. Kasarskis, “A window into third-generation sequencing,” *Human Molecular Genetics*, vol. 19, pp. R227–240, Oct. 2010.

[50] L. Liu, Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu, and M. Law, “Comparison of next-generation sequencing systems,” *Journal of Biomedicine & Biotechnology*, vol. 2012, p. 251364, 2012.

[51] A. Rhoads and K. F. Au, “PacBio Sequencing and Its Applications,” *Genomics, Proteomics & Bioinformatics*, vol. 13, pp. 278–289, Oct. 2015.

[52] L. D. Stein, “The case for cloud computing in genome informatics,” *Genome Biology*, vol. 11, p. 207, May 2010.

[53] J. B. Hagen, “The origins of bioinformatics,” *Nature Reviews. Genetics*, vol. 1, no. 3, pp. 231–236, 2000.

[54] G. Cochrane, I. Karsch-Mizrachi, T. Takagi, and I. N. Sequence Database Collaboration, “The International Nucleotide Sequence Database Collaboration,” *Nucleic Acids Research*, vol. 44, pp. D48–D50, Jan. 2016.

[55] G. Moore, “Cramming more components onto integrated circuits,” *Electronics*, vol. 38, p. 114 ff., Apr. 1965.

[56] A. O’Driscoll, J. Daugelaite, and R. D. Sleator, “’Big data’, Hadoop and cloud computing in genomics,” *Journal of Biomedical Informatics*, vol. 46, pp. 774–781, Oct. 2013.

[57] W. J. S. Diniz and F. Canduri, “REVIEW-ARTICLE Bioinformatics: an overview and its applications,” *Genetics and molecular research: GMR*, vol. 16, Mar. 2017.

[58] S. B. Needleman and C. D. Wunsch, “A general method applicable to the search for similarities in the amino acid sequence of two proteins,” *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.

[59] T. F. Smith and M. S. Waterman, “Identification of common molecular subsequences,” *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195–197, 1981.

[60] W. R. Pearson, “Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms,” *Genomics*, vol. 11, pp. 635–650, Nov. 1991.

[61] D. J. Lipman and W. R. Pearson, “Rapid and sensitive protein similarity searches,” *Science (New York, N.Y.)*, vol. 227, pp. 1435–1441, Mar. 1985.

[62] W. R. Pearson and D. J. Lipman, “Improved tools for biological sequence comparison.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 85, pp. 2444–2448, Apr. 1988.

[63] W. R. Pearson, “Empirical statistical estimates for sequence similarity searches,” *Journal of Molecular Biology*, vol. 276, pp. 71–84, Feb. 1998.

[64] W. R. Pearson, “Flexible sequence similarity searching with the FASTA3 program package,” *Methods in Molecular Biology (Clifton, N.J.)*, vol. 132, pp. 185–219, 2000.

[65] W. R. Pearson, “Finding Protein and Nucleotide Similarities with FASTA,” *Current protocols in bioinformatics*, vol. 53, pp. 3.9.1–3.925, Mar. 2016.

[66] S. Henikoff and J. G. Henikoff, “Amino acid substitution matrices from protein blocks.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, pp. 10915–10919, Nov. 1992.

[67] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” *Journal of molecular biology*, vol. 215, pp. 403–410, Oct. 1990.

[68] A. Morgulis, G. Coulouris, Y. Raytselis, T. L. Madden, R. Agarwala, and A. A. Schäffer, "Database indexing for production MegaBLAST searches," *Bioinformatics (Oxford, England)*, vol. 24, pp. 1757–1764, Aug. 2008.

[69] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden, "BLAST+: architecture and applications," *BMC bioinformatics*, vol. 10, p. 421, Dec. 2009.

[70] P. D. Vouzis and N. V. Sahinidis, "GPU-BLAST: using graphics processors to accelerate protein sequence alignment," *Bioinformatics*, vol. 27, pp. 182–188, Jan. 2011.

[71] B. Buchfink, C. Xie, and D. H. Huson, "Fast and sensitive protein alignment using DIAMOND," *Nature Methods*, vol. 12, pp. 59–60, Jan. 2015.

[72] V. Novichkov, A. Kaznadzey, N. Alexandrova, and D. Kaznadzey, "NSimScan: DNA comparison tool with increased speed, sensitivity and accuracy," *Bioinformatics (Oxford, England)*, vol. 32, no. 15, pp. 2380–2381, 2016.

[73] K. Reinert, T. H. Dadi, M. Ehrhardt, H. Hauswedell, S. Mehringer, R. Rahn, J. Kim, C. Pockrandt, J. Winkler, E. Siragusa, G. Urgese, and D. Weese, "The SeqAn C++ template library for efficient sequence analysis: A resource for programmers," *Journal of Biotechnology*, vol. 261, pp. 157–168, Nov. 2017.

[74] R. Rahn, S. Budach, P. Costanza, M. Ehrhardt, J. Hancox, and K. Reinert, "Generic accelerated sequence alignment in SeqAn using vectorization and multi-threading," *Bioinformatics (Oxford, England)*, vol. 34, pp. 3437–3445, Oct. 2018.

[75] L. Wang and T. Jiang, "On the Complexity of Multiple Sequence Alignment," *Journal of Computational Biology*, vol. 1, pp. 337–348, Jan. 1994.

[76] C. H. Papadimitriou and M. Yannakakis, "Optimization, approximation, and complexity classes," *Journal of Computer and System Sciences*, vol. 43, no. 3, pp. 425–440, 1991. MSC2010: 68Q15 = Complexity classes of computation MSC2010: 68Q25 = Analysis of algorithms and problem complexity.

[77] K. Katoh, K. Misawa, K.-i. Kuma, and T. Miyata, "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform," *Nucleic Acids Research*, vol. 30, pp. 3059–3066, July 2002.

[78] F. Jeanmougin, J. D. Thompson, M. Gouy, D. G. Higgins, and T. J. Gibson, "Multiple sequence alignment with Clustal X," *Trends in biochemical sciences*, vol. 23, pp. 403–405, Oct. 1998.

[79] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson, and D. G. Higgins, "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega," *Molecular systems biology*, vol. 7, p. 539, 2011.

[80] C. Notredame, D. G. Higgins, and J. Heringa, "T-coffee: a novel method for fast and accurate multiple sequence alignment," *Journal of Molecular Biology*, vol. 302, pp. 205–217, Sept. 2000.

[81] C. Magis, J.-F. Taly, G. Bussotti, J.-M. Chang, P. Di Tommaso, I. Erb, J. Espinosa-Carrasco, and C. Notredame, "T-Coffee: Tree-based consistency objective function for alignment evaluation," *Methods in Molecular Biology (Clifton, N.J.)*, vol. 1079, pp. 117–129, 2014.

[82] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research*, vol. 32, no. 5, pp. 1792–1797, 2004.

[83] R. Durbin, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Apr. 1998.

[84] R. C. Edgar and S. Batzoglou, "Multiple sequence alignment," *Current Opinion in Structural Biology*, vol. 16, pp. 368–373, June 2006.

[85] A. Löytynoja, "Phylogeny-aware alignment with PRANK," *Methods in Molecular Biology (Clifton, N.J.)*, vol. 1079, pp. 155–170, 2014.

[86] B. P. Blackburne and S. Whelan, "Class of Multiple Sequence Alignment Algorithm Affects Genomic Analysis," *Molecular Biology and Evolution*, vol. 30, pp. 642–653, Mar. 2013.

[87] T. J. Wheeler and S. R. Eddy, "nhmmer: DNA homology search with profile HMMs," *Bioinformatics (Oxford, England)*, vol. 29, pp. 2487–2489, Oct. 2013.

[88] L. A. K. Ayad and S. P. Pissis, "MARS: improving multiple circular sequence alignment using refined sequences," *BMC Genomics*, vol. 18, p. 86, 2017.

[89] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nature Methods*, vol. 9, pp. 357–359, Apr. 2012.

[90] H. Li and R. Durbin, "Fast and accurate long-read alignment with Burrows-Wheeler transform," *Bioinformatics (Oxford, England)*, vol. 26, pp. 589–595, Mar. 2010.

[91] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup, "The Sequence Alignment/Map format and SAMtools," *Bioinformatics (Oxford, England)*, vol. 25, pp. 2078–2079, Aug. 2009.

[92] M. J. Chaisson and G. Tesler, "Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory," *BMC bioinformatics*, vol. 13, p. 238, Sept. 2012.

[93] H. Li, "Minimap2: pairwise alignment for nucleotide sequences," *Bioinformatics (Oxford, England)*, vol. 34, pp. 3094–3100, Sept. 2018.

[94] R. C. Edgar, "Search and clustering orders of magnitude faster than BLAST," *Bioinformatics (Oxford, England)*, vol. 26, pp. 2460–2461, Oct. 2010.

[95] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: accelerated for clustering the next-generation sequencing data," *Bioinformatics (Oxford, England)*, vol. 28, pp. 3150–3152, Dec. 2012.

[96] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, “An efficient algorithm for large-scale detection of protein families,” *Nucleic Acids Research*, vol. 30, pp. 1575–1584, Apr. 2002.

[97] S. Van Dongen, “A Cluster algorithm for graphs,” *Report - Information systems*, no. 10, pp. 1–40.

[98] P. A. Gagniuc, *Markov chains: from theory to implementation and experimentation.* Hoboken, NJ: John Wiley & Sons, 2017.

[99] D. R. Smith, “Ligation-mediated PCR of restriction fragments from large DNA molecules,” *PCR methods and applications*, vol. 2, pp. 21–27, Aug. 1992.

[100] M. Schmidt, H. Glimm, M. Wissler, G. Hoffmann, K. Olsson, S. Sellers, D. Carbonaro, J. F. Tisdale, C. Leurs, H. Hanenberg, C. E. Dunbar, H.-P. Kiem, S. Karlsson, D. B. Kohn, D. Williams, and C. Von Kalle, “Efficient characterization of retro-, lenti-, and foamyvector-transduced cell populations by high-accuracy insertion site sequencing,” *Annals of the New York Academy of Sciences*, vol. 996, pp. 112–121, May 2003.

[101] M. Schmidt, K. Schwarzwaelder, C. Bartholomae, K. Zaoui, C. Ball, I. Pilz, S. Braun, H. Glimm, and C. von Kalle, “High-resolution insertion-site analysis by linear amplification-mediated PCR (LAM-PCR),” *Nature Methods*, vol. 4, pp. 1051–1057, Dec. 2007.

[102] R. Gabriel, R. Eckenberg, A. Paruzynski, C. C. Bartholomae, A. Nowrouzi, A. Arens, S. J. Howe, A. Recchia, C. Cattoglio, W. Wang, K. Faber, K. Schwarzwaelder, R. Kirsten, A. Deichmann, C. R. Ball, K. S. Balaggan, R. J. Yáñez-Muñoz, R. R. Ali, H. B. Gaspar, L. Biasco, A. Aiuti, D. Cesana, E. Montini, L. Naldini, O. Cohen-Haguenauer, F. Mavilio, A. J. Thrasher, H. Glimm, C. von Kalle, W. Saurin, and M. Schmidt, “Comprehensive genomic access to vector integration in clinical gene therapy,” *Nature Medicine*, vol. 15, pp. 1431–1436, Dec. 2009.

[103] T. B. Hawkins, J. Dantzer, B. Peters, M. Dinauer, K. Mockaitis, S. Mooney, and K. Cornetta, “Identifying viral integration sites using SeqMap 2.0,” *Bioinformatics*, vol. 27, pp. 720–722, Mar. 2011.

[104] G. O. Sperber, T. Airola, P. Jern, and J. Blomberg, “Automated recognition of retroviral sequences in genomic data—RetroTector©,” *Nucleic Acids Research*, vol. 35, pp. 4964–4976, Aug. 2007.

[105] A. Calabria, S. Leo, F. Benedicenti, D. Cesana, G. Spinozzi, M. Orsini, S. Merella, E. Stupka, G. Zanetti, and E. Montini, “VISPA: a computational pipeline for the identification and analysis of genomic vector integration sites,” *Genome Medicine*, vol. 6, Sept. 2014.

[106] Y. Chen, H. Yao, E. J. Thompson, N. M. Tannir, J. N. Weinstein, and X. Su, “VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue,” *Bioinformatics (Oxford, England)*, vol. 29, pp. 266–267, Jan. 2013.

[107] J.-W. Li, R. Wan, C.-S. Yu, N. N. Co, N. Wong, and T.-F. Chan, "ViralFusionSeq: accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution," *Bioinformatics (Oxford, England)*, vol. 29, pp. 649–651, Mar. 2013.

[108] Q. Wang, P. Jia, and Z. Zhao, "VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data," *PloS One*, vol. 8, no. 5, p. e64465, 2013.

[109] D. W. Ho, K. M. Sze, and I. O. Ng, "Virus-Clip: a fast and memory-efficient viral integration site detection tool at single-base resolution with annotation capability," *Oncotarget*, vol. 6, pp. 20959–20963, May 2015.

[110] A. Kamboj, C. V. Hallwirth, I. E. Alexander, G. B. McCowage, and B. Kramer, "Ub-ISAP: a streamlined UNIX pipeline for mining unique viral vector integration sites from next generation sequencing data," *BMC Bioinformatics*, vol. 18, June 2017.

[111] N.-p. D. Nguyen, V. Deshpande, J. Luebeck, P. S. Mischel, and V. Bafna, "ViFi: accurate detection of viral integration and mRNA fusion reveals indiscriminate and unregulated transcription in proximal genomic regions in cervical cancer," *Nucleic Acids Research*, vol. 46, pp. 3309–3325, Apr. 2018.

[112] S. D. Rivas-Carrillo, M. E. Pettersson, C.-J. Rubin, and P. Jern, "Whole-genome comparison of endogenous retrovirus segregation across wild and domestic host species populations," *Proceedings of the National Academy of Sciences*, vol. 115, pp. 11012–11017, Oct. 2018.

[113] O. Yeku and M. A. Frohman, "Rapid amplification of cDNA ends (RACE)," *Methods in Molecular Biology (Clifton, N.J.)*, vol. 703, pp. 107–122, 2011.

[114] O. S. Kustikova, U. Modlich, and B. Fehse, "Retroviral insertion site analysis in dominant haematopoietic clones," *Methods in Molecular Biology (Clifton, N.J.)*, vol. 506, pp. 373–390, 2009.

[115] D. Hüser, A. Gogol-Döring, T. Lutter, S. Weger, K. Winter, E.-M. Hammer, T. Cathomen, K. Reinert, and R. Heilbronn, "Integration Preferences of Wildtype AAV-2 for Consensus Rep-Binding Sites at Numerous Loci in the Human Genome," *PLoS Pathogens*, vol. 6, July 2010.

[116] F. M. Shapter and D. L. E. Waters, "Genome walking," *Methods in Molecular Biology (Clifton, N.J.)*, vol. 1099, pp. 133–146, 2014.

[117] F. Bushman, M. Lewinski, A. Ciuffi, S. Barr, J. Leipzig, S. Hannenhalli, and C. Hoffmann, "Genome-wide analysis of retroviral DNA integration," *Nature Reviews. Microbiology*, vol. 3, pp. 848–858, Nov. 2005.

[118] J. J. Hanger, L. D. Bromham, J. J. McKee, T. M. O'Brien, and W. F. Robinson, "The nucleotide sequence of koala (Phascolarctos cinereus) retrovirus: a novel type C endogenous virus related to Gibbon ape leukemia virus," *Journal of virology*, vol. 74, pp. 4264–4272, May 2000.

[119] R. E. Tarlinton, J. Meers, and P. R. Young, "Retroviral invasion of the koala genome," *Nature*, vol. 442, pp. 79–81, July 2006.

[120] W. Xu, C. K. Stadler, K. Gorman, N. Jensen, D. Kim, H. Zheng, S. Tang, W. M. Switzer, G. W. Pye, and M. V. Eiden, "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*, vol. 110, pp. 11547–11552, July 2013.

[121] M. C. Ávila Arcos, S. Y. W. Ho, Y. Ishida, N. Nikolaidis, K. Tsangaras, K. Hönig, R. Medina, M. Rasmussen, S. L. Fordyce, S. Calvignac-Spencer, E. Willerslev, M. T. P. Gilbert, K. M. Helgen, A. L. Roca, and A. D. Greenwood, "One hundred twenty years of koala retrovirus evolution determined from museum skins," *Molecular biology and evolution*, vol. 30, pp. 299–304, Feb. 2013.

[122] M. Hobbs, A. Pavasovic, A. G. King, P. J. Prentis, M. D. Eldridge, Z. Chen, D. J. Colgan, A. Polkinghorne, M. R. Wilkins, C. Flanagan, A. Gillett, J. Hanger, R. N. Johnson, and P. Timms, "A transcriptome resource for the koala (Phascolarctos cinereus): insights into koala retrovirus transcription and sequence diversity," *BMC Genomics*, vol. 15, p. 786, Sept. 2014.

[123] K. Tsangaras, N. Wales, T. Sicheritz-Pontén, S. Rasmussen, J. Michaux, Y. Ishida, S. Morand, M.-L. Kampmann, M. T. P. Gilbert, and A. D. Greenwood, "Hybridization capture using short PCR products enriches small genomes by capturing flanking sequences (CapFlank)," *PloS One*, vol. 9, no. 10, p. e109101, 2014.

[124] N. Alfano, J. Michaux, S. Morand, K. Aplin, K. Tsangaras, U. Löber, P.-H. Fabre, Y. Fitriana, G. Semiadi, Y. Ishida, K. M. Helgen, A. L. Roca, M. V. Eiden, and A. D. Greenwood, "Endogenous Gibbon Ape Leukemia Virus Identified in a Rodent (Melomys burtoni subsp.) from Wallacea (Indonesia)," *Journal of Virology*, vol. 90, pp. 8169–8180, Sept. 2016.

[125] E. Cappellini, A. Prohaska, F. Racimo, F. Welker, M. W. Pedersen, M. E. Allentoft, P. de Barros Damgaard, P. Gutenbrunner, J. Dunne, S. Hammann, M. Roffet-Salque, M. Ilardo, J. V. Moreno-Mayar, Y. Wang, M. Sikora, L. Vinner, J. Cox, R. P. Evershed, and E. Willerslev, "Ancient Biomolecules and Evolutionary Inference," *Annual Review of Biochemistry*, vol. 87, no. 1, pp. 1029–1060, 2018.

[126] A. Nowrouzi, M. Dittrich, C. Klanke, M. Heinkelein, M. Rammling, T. Dandekar, C. von Kalle, and A. Rethwilm, "Genome-wide mapping of foamy virus vector integrations into a human cell line," *The Journal of General Virology*, vol. 87, pp. 1339–1347, May 2006.

[127] Y. Moalic, Y. Blanchard, H. Félix, and A. Jestin, "Porcine Endogenous Retrovirus Integration Sites in the Human Genome: Features in Common with Those of Murine Leukemia Virus," *Journal of Virology*, vol. 80, pp. 10980–10988, Nov. 2006.

[128] C. S. Schmidt, K. A. Hultman, D. Robinson, K. Killham, and J. I. Prosser, "PCR profiling of ammonia-oxidizer communities in acidic soils subjected to nitrogen and sulphur deposition," *FEMS microbiology ecology*, vol. 61, pp. 305–316, Aug. 2007.

[129] A. Ciuffi and S. D. Barr, "Identification of HIV integration sites in infected host genomic DNA," *Methods (San Diego, Calif.)*, vol. 53, pp. 39–46, Jan. 2011.

[130] P. Brotherton, P. Endicott, M. Beaumont, R. Barnett, J. Austin, A. Cooper, and J. J. Sanchez, "Single primer extension (SPEX) amplification to accurately genotype highly damaged DNA templates," *Forensic Science International: Genetics Supplement Series*, vol. 1, pp. 19–21, Aug. 2008.

[131] A. W. Briggs, J. M. Good, R. E. Green, J. Krause, T. Maricic, U. Stenzel, and S. Pääbo, "Primer extension capture: targeted sequence retrieval from heavily degraded DNA sources," *Journal of Visualized Experiments: JoVE*, no. 31, p. 1573, 2009.

[132] T. Maricic, M. Whitten, and S. Pääbo, "Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products," *PLoS ONE*, vol. 5, p. e14004, Nov. 2010.

[133] G. Sperber, A. Lövgren, N.-E. Eriksson, F. Benachenhou, and J. Blomberg, "RetroTector online, a rational tool for analysis of retroviral elements in small and medium size vertebrate genomic sequences," *BMC Bioinformatics*, vol. 10, p. S4, June 2009.

[134] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," *EMBnet.journal*, vol. 17, pp. pp. 10–12, May 2011.

[135] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: A flexible trimmer for Illumina Sequence Data," *Bioinformatics*, p. btu170, Apr. 2014.

[136] T. Magoč and S. L. Salzberg, "FLASH: Fast Length Adjustment of Short Reads to Improve Genome Assemblies," *Bioinformatics*, p. btr507, Sept. 2011.

[137] P. Rice, I. Longden, and A. Bleasby, "EMBOSS: the European Molecular Biology Open Software Suite," *Trends in genetics: TIG*, vol. 16, pp. 276–277, June 2000.

[138] Z. Zhang, S. Schwartz, L. Wagner, and W. Miller, "A greedy algorithm for aligning DNA sequences," *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, vol. 7, pp. 203–214, Apr. 2000.

[139] K. Katoh and D. M. Standley, "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability," *Molecular Biology and Evolution*, vol. 30, pp. 772–780, Apr. 2013.

[140] A. M. Waterhouse, J. B. Procter, D. M. A. Martin, M. Clamp, and G. J. Barton, "Jalview Version 2–a multiple sequence alignment editor and analysis workbench," *Bioinformatics (Oxford, England)*, vol. 25, pp. 1189–1191, May 2009.

[141] Y. Ishida, C. McCallister, N. Nikolaidis, K. Tsangaras, K. M. Helgen, A. D. Greenwood, and A. L. Roca, "Sequence variation of koala retrovirus transmembrane protein p15e among koalas from different geographic regions," *Virology*, vol. 475, pp. 28–36, Jan. 2015.

[142] C. J. Mulligan, "Isolation and analysis of DNA from archaeological, clinical, and natural history specimens," *Methods in Enzymology*, vol. 395, pp. 87–103, 2005.

[143] M. Knapp and M. Hofreiter, "Next Generation Sequencing of Ancient DNA: Requirements, Strategies and Perspectives," *Genes*, vol. 1, pp. 227–243, July 2010.

[144] A. Döring, D. Weese, T. Rausch, and K. Reinert, "SeqAn An efficient, generic C++ library for sequence analysis," *BMC Bioinformatics*, vol. 9, p. 11, Jan. 2008.

[145] B. T. James, B. B. Luczak, and H. Z. Girgis, "MeShClust: an intelligent tool for clustering DNA sequences," *Nucleic Acids Research*, vol. 46, pp. e83–e83, Aug. 2018.

[146] R. N. Johnson, D. O'Meally, Z. Chen, G. J. Etherington, S. Y. W. Ho, W. J. Nash, C. E. Grueber, Y. Cheng, C. M. Whittington, S. Dennison, E. Peel, W. Haerty, R. J. O'Neill, D. Colgan, T. L. Russell, D. E. Alquezar-Planas, V. Attenbrow, J. G. Bragg, P. A. Brandies, A. Y.-Y. Chong, J. E. Deakin, F. D. Palma, Z. Duda, M. D. B. Eldridge, K. M. Ewart, C. J. Hogg, G. J. Frankham, A. Georges, A. K. Gillett, M. Govendir, A. D. Greenwood, T. Hayakawa, K. M. Helgen, M. Hobbs, C. E. Holleley, T. N. Heider, E. A. Jones, A. King, D. Madden, J. A. M. Graves, K. M. Morris, L. E. Neaves, H. R. Patel, A. Polkinghorne, M. B. Renfree, C. Robin, R. Salinas, K. Tsangaras, P. D. Waters, S. A. Waters, B. Wright, M. R. Wilkins, P. Timms, and K. Belov, "Adaptation and conservation insights from the koala genome," *Nature Genetics*, vol. 50, pp. 1102–1111, Aug. 2018.

[147] J. E. Vargas, L. Chicaybam, R. T. Stein, A. Tanuri, A. Delgado-Cañedo, and M. H. Bonamino, "Retroviral vectors and transposons for stable gene therapy: advances, current challenges and perspectives," *Journal of Translational Medicine*, vol. 14, no. 1, p. 288, 2016.

[148] S. Hacein-Bey-Abina, C. von Kalle, M. Schmidt, F. Le Deist, N. Wulffraat, E. McIntyre, I. Radford, J.-L. Villeval, C. C. Fraser, M. Cavazzana-Calvo, and A. Fischer, "A serious adverse event after successful gene therapy for X-linked severe combined immunodeficiency," *The New England Journal of Medicine*, vol. 348, pp. 255–256, Jan. 2003.

[149] M. Schmidt, P. Zickler, G. Hoffmann, S. Haas, M. Wissler, A. Muessig, J. F. Tisdale, K. Kuramoto, R. G. Andrews, T. Wu, H.-P. Kiem, C. E. Dunbar, and C. von Kalle, "Polyclonal long-term repopulating stem cell clones in a primate model," *Blood*, vol. 100, pp. 2737–2743, Oct. 2002.

[150] P. R. Mueller and B. Wold, "In vivo footprinting of a muscle specific enhancer by ligation mediated PCR," *Science (New York, N.Y.)*, vol. 246, pp. 780–786, Nov. 1989.

[151] R. S. Devon, D. J. Porteous, and A. J. Brookes, "Splinkerettes–improved vectorettes for greater efficiency in PCR walking.," *Nucleic Acids Research*, vol. 23, pp. 1644–1645, May 1995.

[152] F. A. Giordano, J.-U. Appelt, B. Link, S. Gerdes, C. Lehrer, S. Scholz, A. Paruzynski, I. Roeder, F. Wenz, H. Glimm, C. von Kalle, M. Grez, M. Schmidt, and S. Laufs, "High-throughput monitoring of integration site clonality in preclinical and clinical gene therapy studies," *Molecular Therapy. Methods & Clinical Development*, vol. 2, p. 14061, 2015.

[153] P. Brotherton, P. Endicott, J. J. Sanchez, M. Beaumont, R. Barnett, J. Austin, and A. Cooper, "Novel high-resolution characterization of ancient DNA reveals C

> U-type base modification events as the sole cause of post mortem miscoding lesions," *Nucleic Acids Research*, vol. 35, no. 17, pp. 5717–5728, 2007.

[154] D. Ustek, S. Sirma, E. Gumus, M. Arikan, A. Cakiris, N. Abaci, J. Mathew, Z. Emrence, H. Azakli, F. Cosan, A. Cakar, M. Parlak, and O. Kursun, "A genome-wide analysis of lentivector integration sites using targeted sequence capture and next generation sequencing technology," *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, vol. 12, pp. 1349–1354, Oct. 2012.

[155] E. J. Duncavage, V. Magrini, N. Becker, J. R. Armstrong, R. T. Demeter, T. Wylie, H. J. Abel, and J. D. Pfeifer, "Hybrid Capture and Next-Generation Sequencing Identify Viral Integration Sites from Formalin-Fixed, Paraffin-Embedded Tissue," *The Journal of Molecular Diagnostics : JMD*, vol. 13, pp. 325–333, May 2011.

[156] T. Karamitros and G. Magiorkinis, "A novel method for the multiplexed target enrichment of MinION next generation sequencing libraries using PCR-generated baits," *Nucleic Acids Research*, vol. 43, p. e152, Dec. 2015.

[157] J. Dapprich, D. Ferriola, K. Mackiewicz, P. M. Clark, E. Rappaport, M. D'Arcy, A. Sasson, X. Gai, J. Schug, K. H. Kaestner, and D. Monos, "The next generation of target capture technologies - large DNA fragment enrichment and sequencing determines regional genomic variation of high complexity," *BMC genomics*, vol. 17, p. 486, 2016.

[158] M. Giolai, P. Paajanen, W. Verweij, L. Percival-Alwyn, D. Baker, K. Witek, F. Jupe, G. Bryan, I. Hein, J. D. G. Jones, and M. D. Clark, "Targeted capture and sequencing of gene-sized DNA molecules," *BioTechniques*, vol. 61, no. 6, pp. 315–322, 2016.

[159] H. Ochman, A. S. Gerber, and D. L. Hartl, "Genetic applications of an inverse polymerase chain reaction," *Genetics*, vol. 120, pp. 621–623, Nov. 1988.

[160] J. Silver and V. Keerikatte, "Novel use of polymerase chain reaction to amplify cellular DNA adjacent to an integrated provirus," *Journal of Virology*, vol. 63, pp. 1924–1928, May 1989.

[161] M. Meyer and M. Kircher, "Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing," *Cold Spring Harbor Protocols*, vol. 2010, p. pdb.prot5448, June 2010.

[162] M. Kearse, R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz, C. Duran, T. Thierer, B. Ashton, P. Meintjes, and A. Drummond, "Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data," *Bioinformatics (Oxford, England)*, vol. 28, pp. 1647–1649, June 2012.

[163] A. Untergasser, I. Cutcutache, T. Koressaar, J. Ye, B. C. Faircloth, M. Remm, and S. G. Rozen, "Primer3–new capabilities and interfaces," *Nucleic Acids Research*, vol. 40, p. e115, Aug. 2012.

[164] W. Sun, X. You, A. Gogol-Döring, H. He, Y. Kise, M. Sohn, T. Chen, A. Klebes, D. Schmucker, and W. Chen, "Ultra-deep profiling of alternatively spliced Drosophila Dscam isoforms by circularization-assisted multi-segment sequencing," *The EMBO journal*, vol. 32, pp. 2029–2038, July 2013.

[165] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Goodwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg, "Genome Sequencing in Open Microfabricated High Density Picoliter Reactors," *Nature*, vol. 437, pp. 376–380, Sept. 2005.

[166] M.-T. Gansauge, T. Gerber, I. Glocke, P. Korlevic, L. Lippik, S. Nagel, L. M. Riehl, A. Schmidt, and M. Meyer, "Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase," *Nucleic Acids Research*, vol. 45, p. e79, June 2017.

[167] B. D. Ondov, N. H. Bergman, and A. M. Phillippy, "Interactive metagenomic visualization in a Web browser," *BMC bioinformatics*, vol. 12, p. 385, Sept. 2011.

[168] M. B. Renfree, A. T. Papenfuss, J. E. Deakin, J. Lindsay, T. Heider, K. Belov, W. Rens, P. D. Waters, E. A. Pharo, G. Shaw, E. S. W. Wong, C. M. Lefèvre, K. R. Nicholas, Y. Kuroki, M. J. Wakefield, K. R. Zenger, C. Wang, M. Ferguson-Smith, F. W. Nicholas, D. Hickford, H. Yu, K. R. Short, H. V. Siddle, S. R. Frankenberg, K. Y. Chew, B. R. Menzies, J. M. Stringer, S. Suzuki, T. A. Hore, M. L. Delbridge, H. R. Patel, A. Mohammadi, N. Y. Schneider, Y. Hu, W. O'Hara, S. Al Nadaf, C. Wu, Z.-P. Feng, B. G. Cocks, J. Wang, P. Flicek, S. M. J. Searle, S. Fairley, K. Beal, J. Herrero, D. M. Carone, Y. Suzuki, S. Sugano, A. Toyoda, Y. Sakaki, S. Kondo, Y. Nishida, S. Tatsumoto, I. Mandiou, A. Hsu, K. A. McColl, B. Lansdell, G. Weinstock, E. Kuczek, A. McGrath, P. Wilson, A. Men, M. Hazar-Rethinam, A. Hall, J. Davis, D. Wood, S. Williams, Y. Sundaravadanam, D. M. Muzny, S. N. Jhangiani, L. R. Lewis, M. B. Morgan, G. O. Okwuonu, S. J. Ruiz, J. Santibanez, L. Nazareth, A. Cree, G. Fowler, C. L. Kovar, H. H. Dinh, V. Joshi, C. Jing, F. Lara, R. Thornton, L. Chen, J. Deng, Y. Liu, J. Y. Shen, X.-Z. Song, J. Edson, C. Troon, D. Thomas, A. Stephens, L. Yapa, T. Levchenko, R. A. Gibbs, D. W. Cooper, T. P. Speed, A. Fujiyama, J. A. M. Graves, R. J. O'Neill, A. J. Pask, S. M. Forrest, and K. C. Worley, "Genome sequence of an Australian kangaroo, Macropus eugenii, provides insight into the evolution of mammalian reproduction and development," *Genome Biology*, vol. 12, no. 8, p. R81, 2011.

[169] L. Aigrain, Y. Gu, and M. A. Quail, "Quantitation of next generation sequencing library preparation protocol efficiencies using droplet digital PCR assays - a systematic comparison of DNA library preparation kits for Illumina sequencing," *BMC genomics*, vol. 17, p. 458, 2016.

[170] M.-T. Gansauge and M. Meyer, “Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA,” *Nature Protocols*, vol. 8, pp. 737–748, Apr. 2013.

[171] M. J. P. Chaisson, J. Huddleston, M. Y. Dennis, P. H. Sudmant, M. Malig, F. Hormozdiari, F. Antonacci, U. Surti, R. Sandstrom, M. Boitano, J. M. Landolin, J. A. Stamatoyannopoulos, M. W. Hunkapiller, J. Korlach, and E. E. Eichler, “Resolving the complexity of the human genome using single-molecule sequencing,” *Nature*, vol. 517, pp. 608–611, Jan. 2015.

[172] M. Jain, S. Koren, K. H. Miga, J. Quick, A. C. Rand, T. A. Sasani, J. R. Tyson, A. D. Beggs, A. T. Dilthey, I. T. Fiddes, S. Malla, H. Marriott, T. Nieto, J. O’Grady, H. E. Olsen, B. S. Pedersen, A. Rhie, H. Richardson, A. R. Quinlan, T. P. Snutch, L. Tee, B. Paten, A. M. Phillippy, J. T. Simpson, N. J. Loman, and M. Loose, “Nanopore sequencing and assembly of a human genome with ultra-long reads,” *Nature Biotechnology*, vol. 36, pp. 338–345, Apr. 2018.

[173] M. Wang, C. R. Beck, A. C. English, Q. Meng, C. Buhay, Y. Han, H. V. Doddapaneni, F. Yu, E. Boerwinkle, J. R. Lupski, D. M. Muzny, and R. A. Gibbs, “PacBio-LITS: a large-insert targeted sequencing method for characterization of human disease-associated chromosomal structural variations,” *BMC genomics*, vol. 16, p. 214, Mar. 2015.

[174] J. D. Merker, A. M. Wenger, T. Sneddon, M. Grove, Z. Zappala, L. Fresard, D. Waggott, S. Utiramerur, Y. Hou, K. S. Smith, S. B. Montgomery, M. Wheeler, J. G. Buchan, C. C. Lambert, K. S. Eng, L. Hickey, J. Korlach, J. Ford, and E. A. Ashley, “Long-read genome sequencing identifies causal structural variation in a Mendelian disease,” *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, vol. 20, pp. 159–163, Jan. 2018.

[175] B. Osborne and P. Schattner, “BioPerl::AlignIO,” 2012.

[176] S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, and A. M. Phillippy, “Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation,” *Genome Research*, vol. 27, pp. 722–736, May 2017.

[177] A. Smit, R. Hubley, and P. Green, “RepeatMasker Open-4.0.,” 2007.

[178] D. R. Zerbino and E. Birney, “Velvet: Algorithms for De Novo Short Read Assembly Using De Bruijn Graphs,” *Genome Research*, vol. 18, pp. 821–829, May 2008.

[179] D. R. Zerbino, “Using the Velvet de novo assembler for short-read sequencing technologies,” *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, vol. CHAPTER, pp. Unit–11.5, Sept. 2010.

[180] T. Shojima, S. Hoshino, M. Abe, J. Yasuda, H. Shogen, T. Kobayashi, and T. Miyazawa, “Construction and characterization of an infectious molecular clone of Koala retrovirus,” *Journal of virology*, vol. 87, pp. 5081–5088, May 2013.

[181] J. U. Pontius, J. C. Mullikin, D. R. Smith, K. Lindblad-Toh, S. Gnerre, M. Clamp, J. Chang, R. Stephens, B. Neelam, N. Volfovsky, A. A. Schäffer, R. Agarwala,

K. Narfström, W. J. Murphy, U. Giger, A. L. Roca, A. Antunes, M. Menotti-Raymond, N. Yuhki, J. Pecon-Slattery, W. E. Johnson, G. Bourque, G. Tesler, and S. J. O'Brien, "Initial sequence and comparative analysis of the cat genome," *Genome Research*, vol. 17, pp. 1675–1689, Nov. 2007.

[182] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowki, and International Human Genome Sequencing Consortium, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860–921, Feb. 2001.

[183] L. Bromham, “The human zoo: endogenous retroviruses in the human genome,” *Trends in Ecology & Evolution*, vol. 17, pp. 91–97, Jan. 2002.

[184] M. Suntsova, A. Garazha, A. Ivanova, D. Kaminsky, A. Zhavoronkov, and A. Buzdin, “Molecular functions of human endogenous retroviruses in health and disease,” *Cellular and molecular life sciences: CMLS*, vol. 72, pp. 3653–3675, Oct. 2015.

[185] V. Blikstad, F. Benachenhou, G. O. Sperber, and J. Blomberg, “Evolution of human endogenous retroviral sequences: a conceptual account,” *Cellular and molecular life sciences: CMLS*, vol. 65, pp. 3348–3365, Nov. 2008.

[186] A. Hayward, M. Grabherr, and P. Jern, “Broad-scale phylogenomics provides insights into retrovirus–host evolution,” *Proceedings of the National Academy of Sciences*, vol. 110, pp. 20146–20151, Dec. 2013.

[187] J. Denner, “Transspecies transmissions of retroviruses: new cases,” *Virology*, vol. 369, pp. 229–233, Dec. 2007.

[188] E. C. Holmes, “The evolution of endogenous viral elements,” *Cell Host & Microbe*, vol. 10, pp. 368–377, Oct. 2011.

[189] M. Escalera-Zamudio and A. D. Greenwood, “On the classification and evolution of endogenous retrovirus: human endogenous retroviruses may not be ‘human’ after all,” *APMIS*, vol. 124, pp. 44–51, Jan. 2016.

[190] N. M. Oliveira, H. Satija, I. A. Kouwenhoven, and M. V. Eiden, “Changes in viral protein function that accompany retroviral endogenization,” *Proceedings of the National Academy of Sciences*, vol. 104, pp. 17506–17511, Oct. 2007.

[191] J. P. Stoye, “Koala retrovirus: a genome invasion in real time,” *Genome Biology*, vol. 7, no. 11, p. 241, 2006.

[192] G. Magiorkinis, R. J. Gifford, A. Katzourakis, J. D. Ranter, and R. Belshaw, “Env-less endogenous retroviruses are genomic superspreaders,” *Proceedings of the National Academy of Sciences*, p. 201200913, Apr. 2012.

[193] L. H. Evans, A. S. M. Alamgir, N. Owens, N. Weber, K. Virtaneva, K. Barbian, A. Babar, F. Malik, and K. Rosenke, “Mobilization of Endogenous Retroviruses in Mice after Infection with an Exogenous Retrovirus,” *Journal of Virology*, vol. 83, pp. 2429–2435, Mar. 2009.

[194] R. Belshaw, J. Watson, A. Katzourakis, A. Howe, J. Woolven-Allen, A. Burt, and M. Tristem, “Rate of Recombinational Deletion among Human Endogenous Retroviruses,” *Journal of Virology*, vol. 81, pp. 9437–9442, Sept. 2007.

[195] A. R. Quinlan and I. M. Hall, “BEDTools: a flexible suite of utilities for comparing genomic features,” *Bioinformatics*, vol. 26, pp. 841–842, Mar. 2010.

[196] H. Li, “Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM,” Mar. 2013.

[197] R. C. Team, “R: A language and environment for statistical computing,” 2013.

[198] E. Paradis, “pegas: an R package for population genetics with an integrated-modular approach,” *Bioinformatics (Oxford, England)*, vol. 26, pp. 419–420, Feb. 2010.

[199] J. Jurka, V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz, “Repbase Update, a database of eukaryotic repetitive elements,” *Cytogenetic and genome research*, vol. 110, no. 1-4, pp. 462–467, 2005.

[200] L. Glover, J. Jun, and D. Horn, “Microhomology-mediated deletion and gene conversion in African trypanosomes,” *Nucleic Acids Research*, vol. 39, pp. 1372–1380, Mar. 2011.

[201] H. Verdin, B. D’haene, D. Beysen, Y. Novikova, B. Menten, T. Sante, P. Lapunzina, J. Nevado, C. M. B. Carvalho, J. R. Lupski, and E. De Baere, “Microhomology-mediated mechanisms underlie non-recurrent disease-causing microdeletions of the FOXL2 gene or its regulatory domain,” *PLoS genetics*, vol. 9, no. 3, p. e1003358, 2013.

[202] L. E. L. M. Vissers, S. S. Bhatt, I. M. Janssen, Z. Xia, S. R. Lalani, R. Pfundt, K. Derwinska, B. B. A. de Vries, C. Gilissen, A. Hoischen, M. Nesteruk, B. Wisniowiecka-Kowalnik, M. Smyk, H. G. Brunner, S. W. Cheung, A. G. van Kessel, J. A. Veltman, and P. Stankiewicz, “Rare pathogenic microdeletions and tandem duplications are microhomology-mediated and stimulated by local genomic architecture,” *Human Molecular Genetics*, vol. 18, pp. 3579–3593, Oct. 2009.

[203] A. Chen, I. T. Weber, R. W. Harrison, and J. Leis, “Identification of amino acids in HIV-1 and avian sarcoma virus integrase subsites required for specific recognition of the long terminal repeat Ends,” *The Journal of Biological Chemistry*, vol. 281, pp. 4173–4182, Feb. 2006.

[204] R. Martin and K. A. Handasyde, *The Koala: Natural History, Conservation and Management.* UNSW Press, 1999. Google-Books-ID: RdWg_f5UI7cC.

[205] A. Flockerzi, S. Burkhardt, W. Schempp, E. Meese, and J. Mayer, “Human Endogenous Retrovirus HERV-K14 Families: Status, Variants, Evolution, and Mobilization of Other Cellular Sequences,” *Journal of Virology*, vol. 79, pp. 2941–2949, Mar. 2005.

[206] D. C. Hancks and H. H. Kazazian, “SVA retrotransposons: Evolution and genetic instability,” *Seminars in Cancer Biology*, vol. 20, pp. 234–245, Aug. 2010.

[207] J. F. Hughes and J. M. Coffin, “Human Endogenous Retroviral Elements as Indicators of Ectopic Recombination Events in the Primate Genome,” *Genetics*, vol. 171, pp. 1183–1194, Nov. 2005.

[208] L. Vargiu, P. Rodriguez-Tomé, G. O. Sperber, M. Cadeddu, N. Grandi, V. Blikstad, E. Tramontano, and J. Blomberg, “Classification and characterization of human endogenous retroviruses; mosaic forms are common,” *Retrovirology*, vol. 13, p. 7, Jan. 2016.

[209] B. A. Houlden, P. R. England, A. C. Taylor, W. D. Greville, and W. B. Sherwin, "Low genetic variability of the koala Phascolarctos cinereus in south-eastern Australia following a severe population bottleneck," *Molecular Ecology*, vol. 5, pp. 269–281, Apr. 1996.

[210] B. A. Houlden, B. H. Costello, D. Sharkey, E. V. Fowler, A. Melzer, W. Ellis, F. Carrick, P. R. Baverstock, and M. S. Elphinstone, "Phylogeographic differentiation in the mitochondrial control region in the koala, Phascolarctos cinereus (Goldfuss 1817)," *Molecular Ecology*, vol. 8, pp. 999–1011, June 1999.

[211] S. Dennison, G. J. Frankham, L. E. Neaves, C. Flanagan, S. FitzGibbon, M. D. B. Eldridge, and R. N. Johnson, "Population genetics of the koala (Phascolarctos cinereus) in north-eastern New South Wales and south-eastern Queensland," *Australian Journal of Zoology*, vol. 64, pp. 402–412, Apr. 2017.

[212] L. E. Neaves, G. J. Frankham, S. Dennison, S. FitzGibbon, C. Flannagan, A. Gillett, E. Hynes, K. Handasyde, K. M. Helgen, K. Tsangaras, A. D. Greenwood, M. D. B. Eldridge, and R. N. Johnson, "Phylogeography of the Koala, (Phascolarctos cinereus), and Harmonising Data to Inform Conservation," *PLOS ONE*, vol. 11, p. e0162207, Sept. 2016.

[213] K. E. Lee, J. M. Seddon, S. Johnston, S. I. FitzGibbon, F. Carrick, A. Melzer, F. Bercovitch, and W. Ellis, "Genetic diversity in natural and introduced island populations of koalas in Queensland," *Australian Journal of Zoology*, vol. 60, pp. 303–310, Feb. 2013.

[214] A. Katzourakis, G. Magiorkinis, A. G. Lim, S. Gupta, R. Belshaw, and R. Gifford, "Larger mammalian body size leads to lower retroviral activity," *PLoS pathogens*, vol. 10, p. e1004214, July 2014.

[215] J. L. Goodier, "Restricting retrotransposons: a review," *Mobile DNA*, vol. 7, p. 16, Aug. 2016.

[216] M. Escalera-Zamudio, M. L. Z. Mendoza, F. Heeger, E. Loza-Rubio, E. Rojas-Anaya, M. L. Méndez-Ojeda, B. Taboada, C. J. Mazzoni, C. F. Arias, and A. D. Greenwood, "A Novel Endogenous Betaretrovirus in the Common Vampire Bat (Desmodus rotundus) Suggests Multiple Independent Infection and Cross-Species Transmission Events," *Journal of Virology*, vol. 89, pp. 5180–5184, Feb. 2015.

[217] M. Ferrarini, M. Moretto, J. A. Ward, N. Šurbanovski, V. Stevanović, L. Giongo, R. Viola, D. Cavalieri, R. Velasco, A. Cestaro, and D. J. Sargent, "An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome," *BMC Genomics*, vol. 14, p. 670, Dec. 2013.

[218] S. Ardui, A. Ameur, J. R. Vermeesch, and M. S. Hestand, "Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics," *Nucleic Acids Research*, vol. 46, pp. 2159–2168, Mar. 2018.

[219] "A reference standard for genome biology," *Nature Biotechnology*, vol. 36, p. 1121, 2018.

[220] W. Shao, J. Shan, M. F. Kearney, X. Wu, F. Maldarelli, J. W. Mellors, B. Luke, J. M. Coffin, and S. H. Hughes, “Retrovirus Integration Database (RID): a public database for retroviral insertion sites into host genomes,” *Retrovirology*, vol. 13, p. 47, July 2016.

A.1 Supplementary Figures and Tables

Table A.1: KoRV Forward Integration Sites Identified in Bilyarra and Their Genomic Locations Relative to the Koala Reference Genome

Scaffold Number	Retrovirus	Orientation Relative to Scaffold Sequence	Position in Scaffold	Target Site Duplication
000062F-078-01	KoRV	forward	4230	ND
000159F-031-01	KoRV	forward	1095	ND
scaf00004	KoRV	forward	23615392	CTTAG
scaf00006	KoRV	forward	17890343	ATACTGA
scaf00007	KoRV	forward	1797116	GGCC
scaf00016	KoRV	forward	11903436	CTAG
scaf00027	KoRV	forward	2853594	ACCTT
scaf00031	KoRV	forward	6150324	AAGT
scaf00039	KoRV	forward	17332612	GTTC
scaf00041	KoRV	forward	12435187	GTAGT
scaf00043	KoRV	forward	10816237	AGAT
scaf00055	KoRV	forward	9889936	AGTCCT
scaf00058	KoRV	forward	9863928	GATG
scaf00081	KoRV	forward	583341	AGGG
scaf00088	KoRV	forward	1318328	AGAGT
scaf00088	KoRV	forward	2525660	(ACAC)*
scaf00088	KoRV	forward	4799221	ND
scaf00106	KoRV	forward	8525460	TGCCT
scaf00127	KoRV	forward	8654899	GTGG
scaf00137	KoRV	forward	5320687	CCAT(g/t)
scaf00137	KoRV	forward	5426576	AAGC
scaf00137	KoRV	forward	7511603	GTAG
scaf00150	KoRV	forward	3620463	GGAT
scaf00164	KoRV	forward	3258696	GAG
scaf00164	KoRV	forward	3402705	(GTGT)*
scaf00164	KoRV	forward	5264524	TAG
scaf00228	KoRV	forward	1503766	AATAG
scaf00241	KoRV	forward	3345654	CCTG
scaf00273	KoRV	forward	3145264	AGAGG
scaf00322	KoRV	forward	213	GAAGTGA
scaf00354	KoRV	forward	1495511	GAGC
scaf00354	KoRV	forward	1537527	ND
scaf00441	KoRV	forward	282050	ATTC

* TSD within repetitive, low complexity region

Table A.2: KoRV Reverse Integration Sites and Integration Sites of Unknown Orientation Identified in Bilyarra and Their Genomic Locations Relative to the Koala Reference Genome

Scaffold number	Retrovirus	Orientation relative to scaffold sequence	Position in scaffold	Target site duplication
scaf00001	KoRV	reverse	3902862	GTAC
scaf00002	KoRV	reverse	25542095	CCAT
scaf00002	KoRV	reverse	32059851	CTAT
scaf00008	KoRV	reverse	24417033	AGAC
scaf00013	KoRV	reverse	14937508	ND*
scaf00031	KoRV	reverse	17598880	AGTACT
scaf00038	KoRV	reverse	7124654	GTATG
scaf00038	KoRV	reverse	12281933	AGGAG
scaf00040	KoRV	reverse	13892345	CAAAACC
scaf00048	KoRV	reverse	8799477	GAAT
scaf00053	KoRV	reverse	5206411	ATCT
scaf00070	KoRV	reverse	7827832	ATACT
scaf00074	KoRV	reverse	2274509	ATAG
scaf00082	KoRV	reverse	9179129	ATTAC
scaf00094	KoRV	reverse	1633494	GAAA
scaf00098	KoRV	reverse	2205920	GAAA
scaf00107	KoRV	reverse	9030033	AGAGT
scaf00111	KoRV	reverse	5981086	GTAG
scaf00150	KoRV	reverse	2758486	AAAC
scaf00159	KoRV	reverse	14458	AGGC
scaf00164	KoRV	reverse	5577461	GCTC
scaf00241	KoRV	reverse	1620865	AGCAG
scaf00279	KoRV	reverse	1841947	ATTCT
scaf00304	KoRV	reverse	865800	AACT
scaf00310	KoRV	reverse	760620	AGTA
scaf00316	KoRV	reverse	1386589	GTCT
scaf00363	KoRV	reverse	1083497	AGAATT
scaf00491	KoRV	reverse	1539	ATTACT
scaf00634	KoRV	reverse	18365	GCTC
scaf00088	KoRV	ND**	4803899	ACTCC
scaf00097	KoRV	ND**	2135856	ATGA
scaf00166	KoRV	ND**	541779	AACAC
scaf00218	KoRV	ND**	1849618	CAAT

* TSD within repetitive, low complexity region

** 5' and 3' LTR could not be determined, since the orientation of the virus relative to the scaffold was not clear

Table A.3: recKoRV Integration Sites Identified in Bilyarra and Their Genomic Locations Relative to the Koala Reference Genome

Scaffold number	Retrovirus	Orientation relative to scaffold sequence	Position in scaffold	Target site duplication
scaf00021	recKoRV1	forward	16384538	ACACT
scaf00024	recKoRV1	forward	4934798	CTTA
scaf00083	recKoRV1	forward	797972	ACAC
scaf00234	recKoRV1	forward	4233389	AAGAT
scaf00275	recKoRV1	forward	2910435	AGGT
scaf00003	recKoRV1	reverse	3474508	ATAT
scaf00014	recKoRV1	reverse	15030324	ACCT
scaf00037	recKoRV1	reverse	16563123	ATGT
scaf00037	recKoRV1	reverse	17582840	GAAG
scaf00069	recKoRV1	reverse	987368	TTGT
scaf00079	recKoRV1	reverse	1270278	CCTGT
scaf00096	recKoRV1	reverse	1287297	(CTCT)*
scaf00164	recKoRV1	reverse	6007491	CTTTTT
scaf00173	recKoRV1	reverse	557197	GTAT
scaf00002	recKoRV***	forward	23377285	TTAC
scaf00035	recKoRV***	forward	5731424	ATGG
scaf00035	recKoRV***	forward	15145894	CTGT
scaf00061	recKoRV***	forward	4737977	(G)CTAT
scaf00107	recKoRV***	forward	4792775	AGGGCTG
scaf00139	recKoRV***	forward	4264851	AAGAT
scaf00031	recKoRV***	reverse	11457756	CATAAGT
scaf00060	recKoRV***	reverse	13428694	TGCAT
scaf00095	recKoRV***	reverse	9508496	ATAGT
scaf00003	recKoRV***	ND**	10330590	TAAC

* TSD within repetitive, low complexity region

** 5' and 3' LTR could not be determined, since the orientation of the virus relative to the scaffold was not clear

*** KoRV recombinant with PhER, but different breakpoints than recKoRV1. Exact breakpoints could not be determined accurately

Table A.4: Alignment of PCR Sanger Products to Viral Domains, Breakpoints, and Insertion Sites

Primer/Scaffold	**1**	**14**	**21**	**24**	**37a**	**37b**	**69**
IS_F	*env*/ LTR/ IS	IS		IS			IS/ *gag*
IS_R		IS	*env*/ LTR			LTR	*env*/ LTR/ IS
KoRVE30F	*env*/ LTR	LTR/ IS		*env*/ LTR/ IS	LTR	LTR	LTR
KoRVE30R	*env*	PhER/ *env*/ IS	bp17, PhER/ *env*	bp17, PhER/ *env*			bp17, PhER/ *env*
GAG_BP_F			PhER	IS	IS		PhER
GAG_IS_R			LTR	IS/ LTR	IS/ LTR	LTR	LTR
D21R	LTR						
D3F	*env*/ LTR						
KoRVEF15	*env*/ LTR						
natKoRVF	*pol*						
natKoRVR	*gag*/ *pol*						
E15F	*env*/ LTR						
E15R	*gag*						
Primer/Scaffold	**79**	**83**	**96**	**139**	**173**	**275**	
IS_F		LTR/ IS/ *gag*			IS/ LTR	IS	
IS_R	?	(bp17?), *env*/ LTR			IS	IS	
KoRVE30F	LTR	*env*/ LTR				LTR/ IS	
KoRVE30R	bp17, PhER/ *env*	PhER/ LTR				bp17, PhER/ *env*	
GAG_BP_F	PhER	LTR					
GAG_IS_R	LTR	LTR/ *gag*	LTR/ *gag*			IS	
KoRVEF15				*env*/ LTR			

Thirteen different primers were used to amplify sequences for Sanger sequencing to reject or confirm the conformation of recKoRV1 candidates. Fourteen candidates were investigated. For three candidates we could not amplify products (see table 4.1.3). KoRV domains, PhER, corresponding insertion sites (IS), and breakpoints bp12/bp17 were aligned with the Sanger products to classify the results. The sequence length for every product is shown in table A.5.

Table A.5: Length of Sanger Sequenced PCR Products

	IS_F	IS_R	KoRVE30F	KoRVE30R	GAG_BP_F	GAG_IS_R	D21R	D3F	KoRVEF15	natKoRVF	natKoRVR	E15F	E15R
1	881			1064			763	1095	1090	1168	1133	1215	794
14	557	423	670	955									
21		866	569	1027	548	569							
24	793		631	923	515	623							
37a			606		582	612							
37b		657	597			594							
69	854	865	542	915	130	568							
79		913	617	998	485	556							
83	937	925	1096	775	823	688							
96						839							
139									1143				
173	829	94											
275	443	458	365	569	540	682							

13 different primers were used to amplify sequences for Sanger sequencing to reject or confirm the conformation of recKoRV1 candidates. 14 candidates were investigated. For three candidates we could not amplify products (see table 4.1.3). In this table, the sequence length for every Sanger product is reported. KoRV domains, PhER, and breakpoints bp12/bp17 were aligned to the sanger products to classify the results, shown in table A.5.

Figures A.1 to A.11 report the detailed investigation of PCR Sanger sequencing products of recKoRV candidate loci. All PCR products are enumerated for every locus. The panels at the top display alignments of corresponding insertion site sequences and KoRV/PhER domains to the PCR products. Unmatched OCR products were indicated by an ‘X”. In every bottom panel, the expected structure of a recKoRV1 insertion with genomic flanks is shown. Color-coded Sanger sequences are aligned next to the schematic expected structure. The color coding is graded from green (high quality) to red (low quality) for every position based on quality information of ABI score. The average quality of the product and the product length was used to create bars aligned to the reference sequence and indexed by the number corresponding to the product from the top panel. Some figures contain additional information about alignments to recKoRV1 breakpoints.

Figure A.1 shows PCR products conducted from the integration site designated as “scaffold 14”. Four out of six primers amplified products from this locus (cross-reference tables A.4 and A.5). All sequences were of low quality. Nevertheless, both integration sites could be amplified as well as a product ranging from ENV through the LTR into the genomic flank. Additionally, one product was spanning PhER and ENV.

Figure A.2 shows PCR products conducted from the integration site designated as “scaffold 21”. One product (x3) could neither be assigned to any of our references nor matched any sequence of NCBI Genbank. Product 1 completely covered one of the recKoRV1 breakpoints.

Figure A.3 shows PCR products conducted from the integration site designated as “scaffold 24”. One product (x2) could neither be assigned to any of our references nor matched any sequence of NCBI Genbank. Two products (7 and 8) covered bp12, while Product 1 covered bp17. This locus was clearly designated as recKoRV1 insertion.

Figures A.4 and A.4 show PCR products conducted from the integration site designated as “scaffold 37 16563123” and “scaffold 37 17582840”. Initial PCR amplification and Sanger sequencing show no evidence for either of the recKoRV1 breakpoints. Experiments conducted later (data not shown) confirmed that both recKoRV1 was integrated into both loci.

Figure A.6 shows PCR products conducted from the integration site designated as "scaffold 69". This was the only case where all primers used amplified products from this locus (cross-reference tables A.4 and A.5). All sequences were of low quality. Bp17 could be confirmed by product 1.

Figure A.7 shows PCR products conducted from the integration site designated as "scaffold 79". Even all sequences were of low quality, product 1 could confirm the presence of bp17 for that locus. As observed for other recKoRV loci discussed, product 3 (integration site primer) amplified a product which could not be assigned to any reference used in this study or NCBI Genbank.

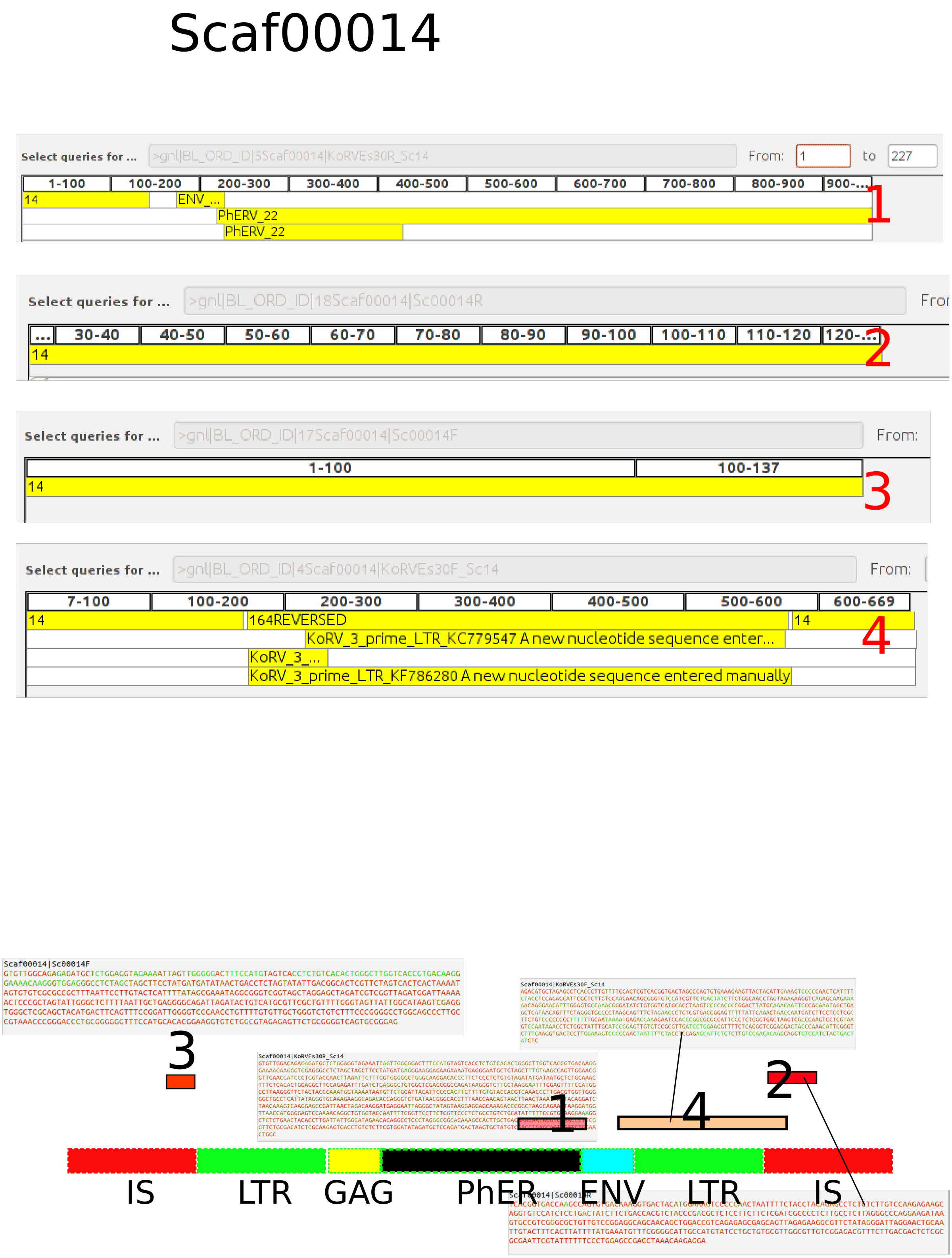
Figure A.8 shows PCR products conducted from the integration site designated as "scaffold 83". Bp17 could be confirmed by the PCR products, while some similarities to bp12 could be found in PCR product 7.

Figure A.9 shows PCR products conducted from the integration site designated as "scaffold 96". Only two out of eight PCR products could be assigned to any reference sequence. Due to the fact that neither PCR amplification and Sanger sequencing, nor mapping of reads from the corresponding sequence cluster to the recKoRV1 breakpoints could validate a recKoRV1 insertion, this integration site was classified as unknown recombinant.

Figure A.10 shows PCR products conducted from the integration site designated as "scaffold 173". Only two out of eight PCR products could be assigned to any reference sequence. Since neither PCR amplification and Sanger sequencing, nor mapping of reads from the corresponding sequence cluster to the recKoRV1 breakpoints could validate a recKoRV1 insertion, this integration site was classified as unknown recombinant.

Figure A.11 shows PCR products conducted from the integration site designated as "scaffold 275". PCR products confirmed bp17. Combined with mapping results of sequences assigned to the sequence cluster from that locus, this locus was classified as recKoRV1 integration site.

Figure A.1: Scaf14 PCR Products Sanger Sequencing


Scaf00014
Select queries for ... >gnl|BL_ORD_ID|5Scaf00014|KoRVEs30R_Sc14 From: 1 to 227
1-100 100-200 200-300 300-400 400-500 500-600 600-700 700-800 800-900 900-...
14
ENV_...
PhERV_22
PhERV_22
1
Select queries for ... >gnl|BL_ORD_ID|18Scaf00014|Sc00014R Fro
... 30-40 40-50 50-60 60-70 70-80 80-90 90-100 100-110 110-120 120-...
14
2
Select queries for ... >gnl|BL_ORD_ID|17Scaf00014|Sc00014F From:
1-100 100-137
14
3
Select queries for ... >gnl|BL_ORD_ID|4Scaf00014|KoRVEs30F_Sc14 From:
7-100 100-200 200-300 300-400 400-500 500-600 600-669
14
164REVERSED
14
KoRV_3_prime_LTR_KC779547 A new nucleotide sequence enter...
KoRV_3_...
KoRV_3_prime_LTR_KF786280 A new nucleotide sequence entered manually
4
3
1
4
2
IS
LTR
GAG
PhER
ENV
LTR
IS

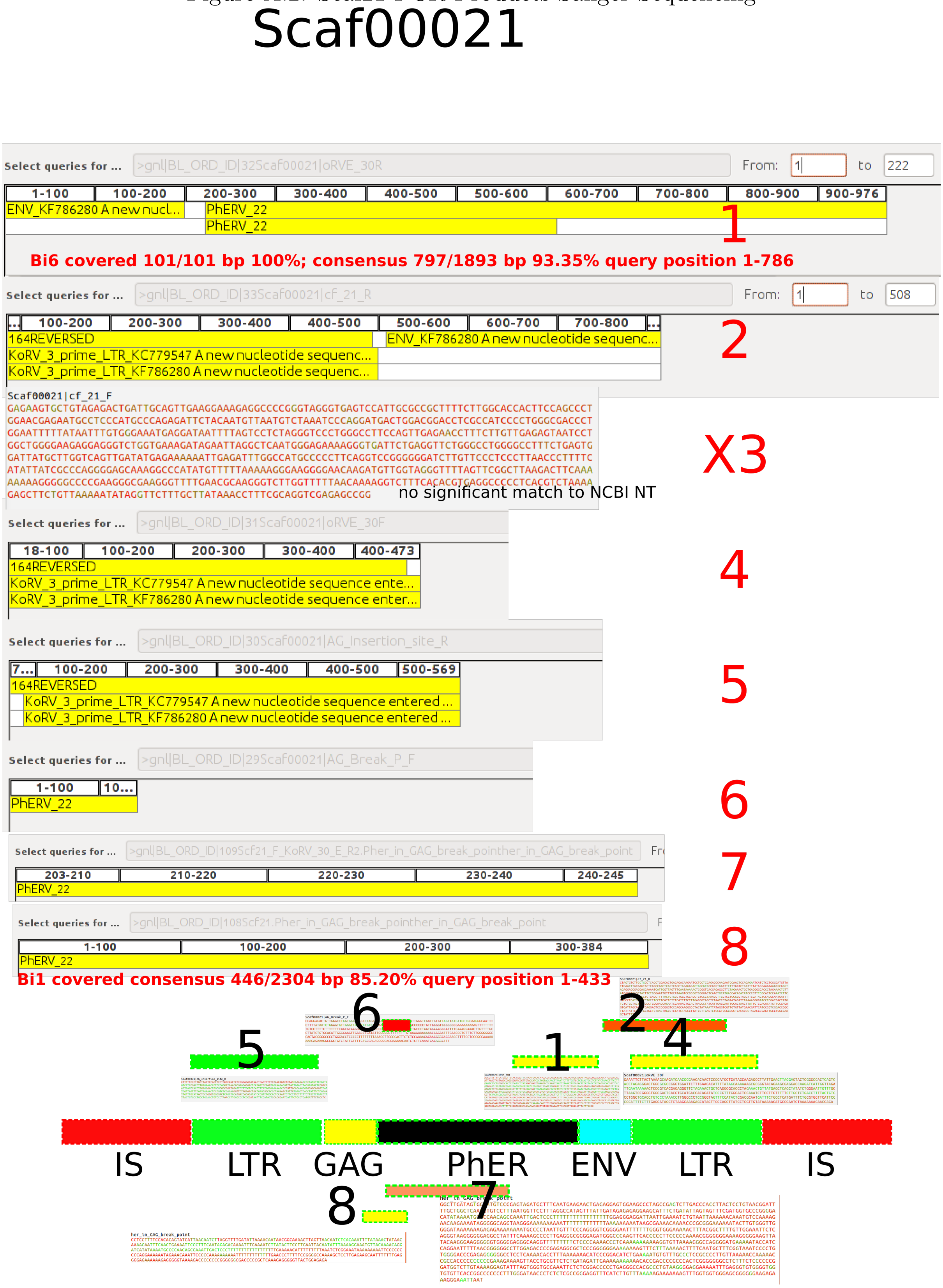
```
Scaf00014|Sc00014F
GTGTTGGCAGAGAGATGCTCTGGAGGTAGAAAATTAGTTGGGGGACTTTCCATGTAGTCACCTCTGTCACACTGGGCTTGGTCACCGTGACAAGG
GAAAACAAGGGTGGAGGGCCTCTAGCTAGCTTCCTATGATGATATAACTGACCTCTAGTATATTGACGGCACTCGTTCTAGTCACTCACTAAAAT
AGTGTGTCGCGCCGCTTTAATTCCTTGTACTCATTTTATAGCGAAATAGGCGGGTCGGTAGCTAGGAGCTAGATCGTCGGTTAGATGGATTAAAA
ACTCCCGCTAGTATTGGGCTCTTTTAATTGCTGAGGGGCAGATTAGATACTGTCATGCGTTCGCTGTTTTGGGTAGTTATTGGCATAAGTCGAGG
TGGGCTCGCAGCTACATGACTTCAGTTTCCGGATTGGGGTCCCAACCTGTTTTGTGTTGCTGGGTCTGTCTTTCCCGGGGCCTGGCAGCCCTTGC
CGTAAACCCGGGACCCTGCGGGGGGTTTCCATGCACACGGAAGGTGTCTGGCGTAGAGAGTTCTGCGGGGTCAGTGCGGGAG
```

```
Scaf00014|KoRVEs30F_Sc14
AGACATGCTAGAGCCTCACCCTTGTTTTCCACTCGTCACGGTGACTAGCCCAGTGTGAAAGAAGTTACTACATTGAAAGTCCCCCAACTCATTTT
CTACCTCCAGAGCATTCGCTCTTGTCCAACAACAGCGGGTGTCCATCGTTCTGACTATCTTCTGGCAACCTAGTAAAAAAGGTCAGAGCAAGAAA
AACAAGGAAGATTTGGAGTGCCAAACGGGATATCTGTGGTCATGCACCTAAGTCCCCACCCCGGACTTATGCAAACAATTCCCAGAAATAGCTGA
GCTCATAACAGTTTCTAGGGTGCCCCTAAGCAGTTTCTAGAACCCTCTCGTGACCGGAGTTTTTATTCAAACTAACCAATGATCTTCCTCCTCGC
TTCTGTCCCCCCCCTTTTTTGCAATAAAATGAGACCAAAGAATCCACCCGGCGCGCCATTCCCTCTGGGTGACTAAGTCGCCCAAGTCCTCGTAA
GTCCAATAAACCTCTGGCTATTTGCATCCGGAGTTGTGTCCGCGTTGATCCTGGAAGGTTTTCTCAGGGTCGGAGGACTACCCAAACATTGGGGT
CTTTCAAGGTGACTCCTTGGAAAGTCCCCCAACTAATTTTCTACCTCCAGAGCATTCTCTCTTGTCCAACACAAGCAGGTGTCCATCTACTGACT
ATCTC
```

```
Scaf00014|KoRVEs30R_Sc14
GTGTTGGACAGAGAGATGCTCTGGAGGTAGAAATTAGTTGGGGGACTTTCCATGTAGTCACCTCTGTCACACTGGGCTTGGTCACCGTGACAAGG
GAAAACAAGGGTGGAGGGCCTCTAGCTAGCTTCCTATGATGAGGGAAGGAGAAGAAAATGAGGGAATGCTGTAGCTTTGTAAGCCAGTTGGAACG
GTTGAACCATCCCTCGTACCAACTTAAATTCTTTGGTGGGGGCTGGGCAAGGACACCCTTCTCCCTCTGTGTAGATATGATAATGCTCTGCAAAC
TTTCTCACACTGGAGGCTTCCAGAGATTTGATCTGAGGGCTGTGGCTCGAGCGGCCAGATAAGGGTCTTGCTAAGGAATTTGGAGTTTTCCATGG
CCTTAAGGGTTCTACTACCCAAATGGTAAAATAATGTTCTGCATTACATTCCCCACTTCTTTTGTGTACCACGTCAAACCCTTGACGTGGTTGGG
GGCTGCCTCATTATAGGGTGCAAAGAAGGCAGACACCAGGGTCTGATAACGGGCACCTTTAACCAACAGTAACTTAACTAAAT[illegible]TACAGGATC
TAACAAAGTCAAGGAGCCGATTAACTAGACAAGGATGAGGAATTAGGGCTATAGTAAGGAGGAGCAAAGACCCGGCTAAGCAGAA[illegible]TAAGGATGG
TTAACCATGGGGAGTCCAAAACAGGCTGTGGTACCAATTTTCGGTTCCTTCTCGTTCCCTCTGCCTGTCTGCATATTTTTCCGTG[illegible]AAGGAAAGG
CTCTCTGAACTACACCTTGATTATTGGCATAGAACACAGGCCTCCCTAGGGCGGCACAAAGCCACTTGCTGAGAAACAATAGGTC[illegible]AACCC[illegible]TCG
GTTCTGCGACATCTCGCAAGAGTGACCTGTCTCTTCGTGGATATAGATGCTCCAGATGACTAAGTGCTATGTC[illegible]ACCTGAGCACT[illegible]TTCTC[illegible]GAA
CTGGC
```

```
Scaf00014|Sc00014R
TCACGGTGACCAAGCCAGTGTGACAAAGGTGACTACATGGAAAGTCCCCCAACTAATTTTCTACCTACAGAGCCTCTCTCTTGTCCAAGAGAAGC
AGGTGTCCATCTCCTGACTATCTTCTGACCACGTCTACCCGACGCTCTCCTTCTTCTCGATCGCCCCTCTTGCCTCTTAGGGCCCAGGAAGATAA
GTGCCGTCGGGCGCTGTTGTCCGGAGGCAGCAACAGCTGGACCGTCAGAGAGCGAGCAGTTAGAGAAGGCGTTCTATAGGGATTAGGAACTGCAA
TTGTACTTTCACTTATTTTATGAAATGTTTCGGGGCATTGCCATGTATCCTGCTGTGCGTTGGCGTTGTCGGAGACGTTTCTTGACGACTCTCGC
GCGAATTCGTATTTTTTCCCTGGAGCCGACCTAAACAAGAGGA
```

Figure A.2: Scaf21 PCR Products Sanger Sequencing



Select queries for ... >gnl|BL_ORD_ID|32Scaf00021|oRVE_30R From: 1 to 222

1-100 | 100-200 | 200-300 | 300-400 | 400-500 | 500-600 | 600-700 | 700-800 | 800-900 | 900-976

ENV_KF786280 A new nucl... PhERV_22

PhERV_22

Bi6 covered 101/101 bp 100%; consensus 797/1893 bp 93.35% query position 1-786

Select queries for ... >gnl|BL_ORD_ID|33Scaf00021|cf_21_R From: 1 to 508

... | 100-200 | 200-300 | 300-400 | 400-500 | 500-600 | 600-700 | 700-800 | ...

164REVERSED ENV_KF786280 A new nucleotide sequenc...

KoRV_3_prime_LTR_KC779547 A new nucleotide sequenc...

KoRV_3_prime_LTR_KF786280 A new nucleotide sequenc...

```
Scaf00021|cf_21_F
GAGAAGTGCTGTAGAGACTGATTGCAGTTGAAGGAAAGAGGCCCCGGGTAGGGTGAGTCCATTGCGCCGCTTTTCTTGGCACCACTTCCAGCCCT
GGAACGAGAATGCCTCCCATGCCCAGAGATTCTACAATGTTAATGTCTAAATCCCAGGATGACTGGACGGACCTCGCCATCCCCTGGGCGACCCT
GGAATTTTTATAATTTGTGGGAAATGAGGATAATTTTAGTCCTCTAGGGTCCCTGGGCCTTCCAGTTGAGAACCTTTCTTGTTGAGAGTAATCCT
GGCTGGGGAAGAGGAGGGTCTGGTGAAAGATAGAATTAGGCTCAATGGGAGAAAAGGGTGATTCTGAGGTTCTGGGCCTGGGGCCTTTCTGAGTG
GATTATGCTTGGTCAGTTGATATGAGAAAAAATTGAGATTTGGCCATGCCCCCTTCAGGTCCGGGGGGATCTTGTTCCCTCCCTTAACCCTTTTC
ATATTATCGCCCAGGGGAGCAAAGGCCCATATGTTTTTAAAAAGGGAAGGGGAACAAGATGTTGGTAGGGTTTTAGTTCGGCTTAAGACTTCAAA
AAAAAGGGGGCCCCGAAGGGCGAAGGGTTTTGAACGCAAGGGTCTTGGTTTTTAACAAAAGGTCTTTCACACGTGAGGCCCCCTCACGTCTAAAA
GAGCTTCTGTTAAAAATATAGGTTCTTTGCTTATAAACCTTTCGCAGGTCGAGAGCCGG
```

no significant match to NCBI NT

Select queries for ... >gnl|BL_ORD_ID|31Scaf00021|oRVE_30F

18-100 | 100-200 | 200-300 | 300-400 | 400-473

164REVERSED

KoRV_3_prime_LTR_KC779547 A new nucleotide sequence ente...

KoRV_3_prime_LTR_KF786280 A new nucleotide sequence enter...

Select queries for ... >gnl|BL_ORD_ID|30Scaf00021|AG_Insertion_site_R

7... | 100-200 | 200-300 | 300-400 | 400-500 | 500-569

164REVERSED

KoRV_3_prime_LTR_KC779547 A new nucleotide sequence entered ...

KoRV_3_prime_LTR_KF786280 A new nucleotide sequence entered ...

Select queries for ... >gnl|BL_ORD_ID|29Scaf00021|AG_Break_P_F

1-100 | 10...

PhERV_22

Select queries for ... >gnl|BL_ORD_ID|109Scf21_F_KoRV_30_E_R2.Pher_in_GAG_break_pointher_in_GAG_break_point Fr

203-210 | 210-220 | 220-230 | 230-240 | 240-245

PhERV_22

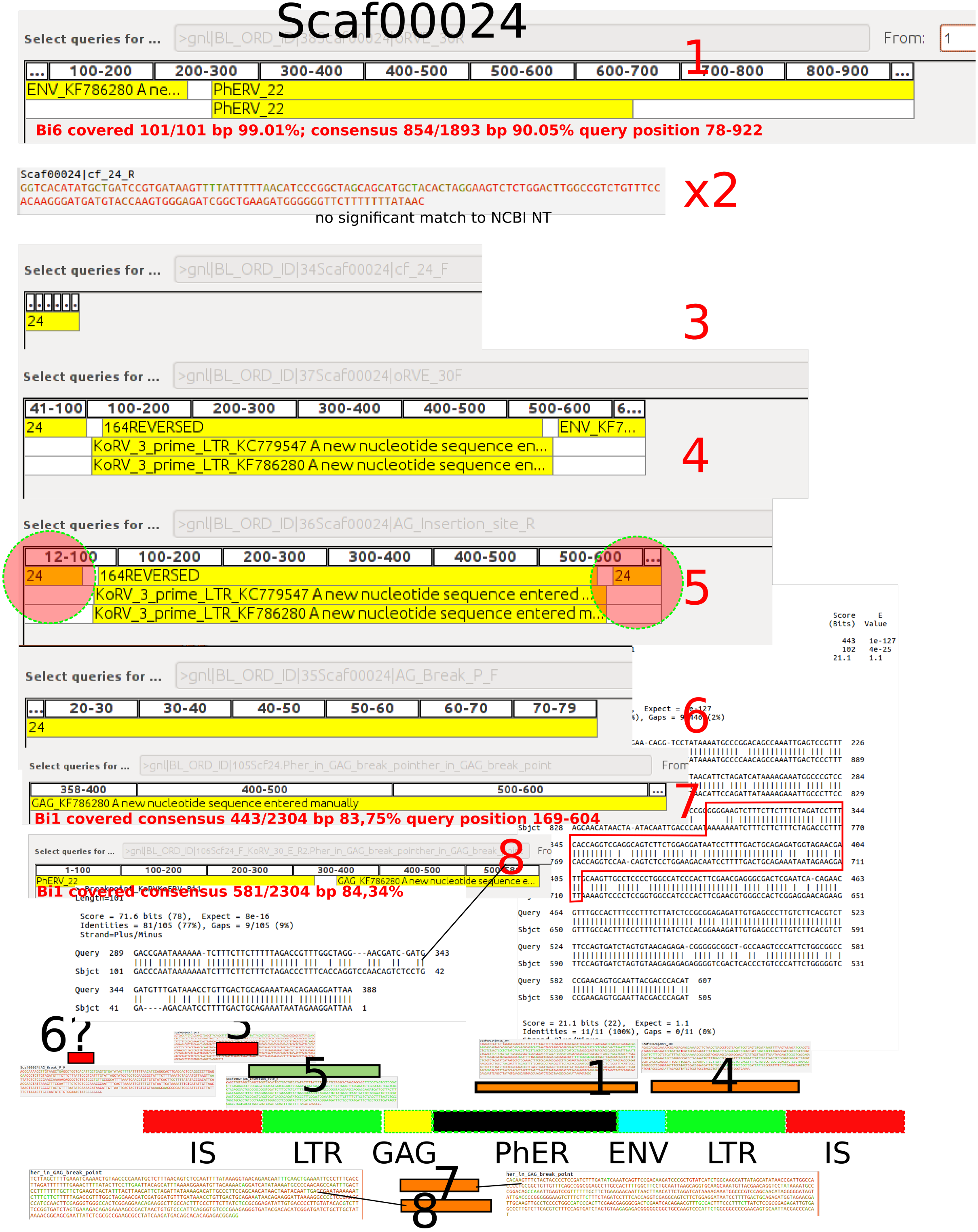
Select queries for ... >gnl|BL_ORD_ID|108Scf21.Pher_in_GAG_break_pointher_in_GAG_break_point F

1-100 | 100-200 | 200-300 | 300-384

PhERV_22

Bi1 covered consensus 446/2304 bp 85.20% query position 1-433

Figure A.3: Scaf24 PCR Products Sanger Sequencing


Scaf00024
Select queries for ...
>gnl|BL_ORD_ID|38Scaf00024|oRVE_30R
From:
1
... 100-200 200-300 300-400 400-500 500-600 600-700 700-800 800-900 ...
ENV_KF786280 A ne...
PhERV_22
PhERV_22
1
x2
Select queries for ...
>gnl|BL_ORD_ID|34Scaf00024|cf_24_F
24
3
Select queries for ...
>gnl|BL_ORD_ID|37Scaf00024|oRVE_30F
41-100 100-200 200-300 300-400 400-500 500-600 6...
24
164REVERSED
ENV_KF7...
KoRV_3_prime_LTR_KC779547 A new nucleotide sequence en...
KoRV_3_prime_LTR_KF786280 A new nucleotide sequence en...
4
Select queries for ...
>gnl|BL_ORD_ID|36Scaf00024|AG_Insertion_site_R
12-100 100-200 200-300 300-400 400-500 500-600 ...
24
164REVERSED
24
KoRV_3_prime_LTR_KC779547 A new nucleotide sequence entered ...
KoRV_3_prime_LTR_KF786280 A new nucleotide sequence entered m...
5
Select queries for ...
>gnl|BL_ORD_ID|35Scaf00024|AG_Break_P_F
... 20-30 30-40 40-50 50-60 60-70 70-79
24
6
Select queries for ...
>gnl|BL_ORD_ID|105Scf24.Pher_in_GAG_break_pointher_in_GAG_break_point
From
358-400 400-500 500-600 ...
GAG_KF786280 A new nucleotide sequence entered manually
7
Select queries for ...
>gnl|BL_ORD_ID|106Scf24_F_KoRV_30_E_R2.Pher_in_GAG_break_pointher_in_GAG_break_point
Fro
1-100 100-200 200-300 300-400 400-500 500-58
PhERV_22
GAG_KF786280 A new nucleotide sequence e...
8
6?
3
5
1
4
IS
LTR
GAG
PhER
ENV
LTR
IS
7
8
Scaf00024|cf_24_F
Scaf00024|AG_Break_P_F
Scaf00024|AG_Insertion_site_R
Scaf00024|oRVE_30R
Scaf00024|oRVE_30F
her_in_GAG_break_point
her_in_GAG_break_point

Bi6 covered 101/101 bp 99.01%; consensus 854/1893 bp 90.05% query position 78-922

```
Scaf00024|cf_24_R
GGTCACATATGCTGATCCGTGATAAGTTTATTTTTAACATCCCGGCTAGCAGCATGCTACACTAGGAAGTCTCTGGACTTGGCCGTCTGTTTCC
ACAAGGGATGATGTACCAAGTGGGAGATCGGCTGAAGATGGGGGGTTCTTTTTTTATAAC
```

no significant match to NCBI NT

Bi1 covered consensus 443/2304 bp 83,75% query position 169-604

Bi1 covered consensus 581/2304 bp 84,34%

```
                                                              Score     E
                                                              (Bits)  Value

                                                               443   1e-127
                                                               102   4e-25
                                                               21.1    1.1

 ,  Expect = 1e-127
%), Gaps = 9/446 (2%)

GAA-CAGG-TCCTATAAAATGCCCGGACAGCCAAATTGAGTCCGTTT  226
          |||||||||||  |||||||||||| ||| |||
          ATAAAATGCCCCAACAGCCAAATTGACTCCCTTT  889

          TAACATTCTAGATCATAAAAGAAATGGCCCGTCC  284
          ||||||| |||| ||||||||||| |||| |||
          AACATTCCAGATTATAAAAGAAATTGCCCTTCC  829

          CCGCGGGGAAGTCTTTCTTCTTTCTAGATCCTTT  344
          |      || |||||||||||||||| |||||
Sbjct  828   AGCAACATAACTA-ATACAATTGACCCAATAAAAAAATCTTTCTTCTTTCTAGACCCTTT  770

       345   CACCAGGTCGAGGCAGTCTTCTGGAGGATAATCCTTTTGACTGCAGAGATGGTAGAACGA  404
             ||||||||| |  |||||| ||||| || ||||||||||||||||| ||  ||||| ||
       769   CACCAGGTCCAA-CAGTCTCCTGGAAGACAATCCTTTTGACTGCAGAAATAATAGAAGGA  711

       405   TTGCAAGTTGCCTCCCCTGGCCATCCCACTTCGAACGAGGGCGACTCGAATCA-CAGAAC  463
             ||  ||||  |||||  |||||||||||||||||||| |||| ||||| |  | |||||
       710   TTAAAAGTCCCCTCCGGTGGCCATCCCACTTCGAACGTGGGCCACTCGGAGGAACAGAAG  651

Query  464   GTTTGCCACTTTCCCTTTCTTATCTCCGCGGAGAGATTGTGAGCCCTTGTCTTCACGTCT  523
             |||||||||||||||||||||||||||| |||| |||||||||||||||||||||||||
Sbjct  650   GTTTGCCACTTTCCCTTTCTTATCTCCACGGAAAGATTGTGAGCCCTTGTCTTCACGTCT  591

Query  524   TTCCAGTGATCTAGTGTAAGAGAGA-CGGGGGCGGCT-GCCAAGTCCCATTCTGGCGGCC  581
             ||||||||||||||||||||||||  |||| || ||  ||  |||||||||||| || |
Sbjct  590   TTCCAGTGATCTAGTGTAAGAGAGAGAGGGGTCGACTCACCCTGTCCCATTCTGGGGGTC  531

Query  582   CCGAACAGTGCAATTACGACCCACAT  607
             ||||| |||| |||||||||||| ||
Sbjct  530   CCGAAGAGTGGAATTACGACCCAGAT  505

 Score = 21.1 bits (22),  Expect = 1.1
 Identities = 11/11 (100%), Gaps = 0/11 (0%)
```

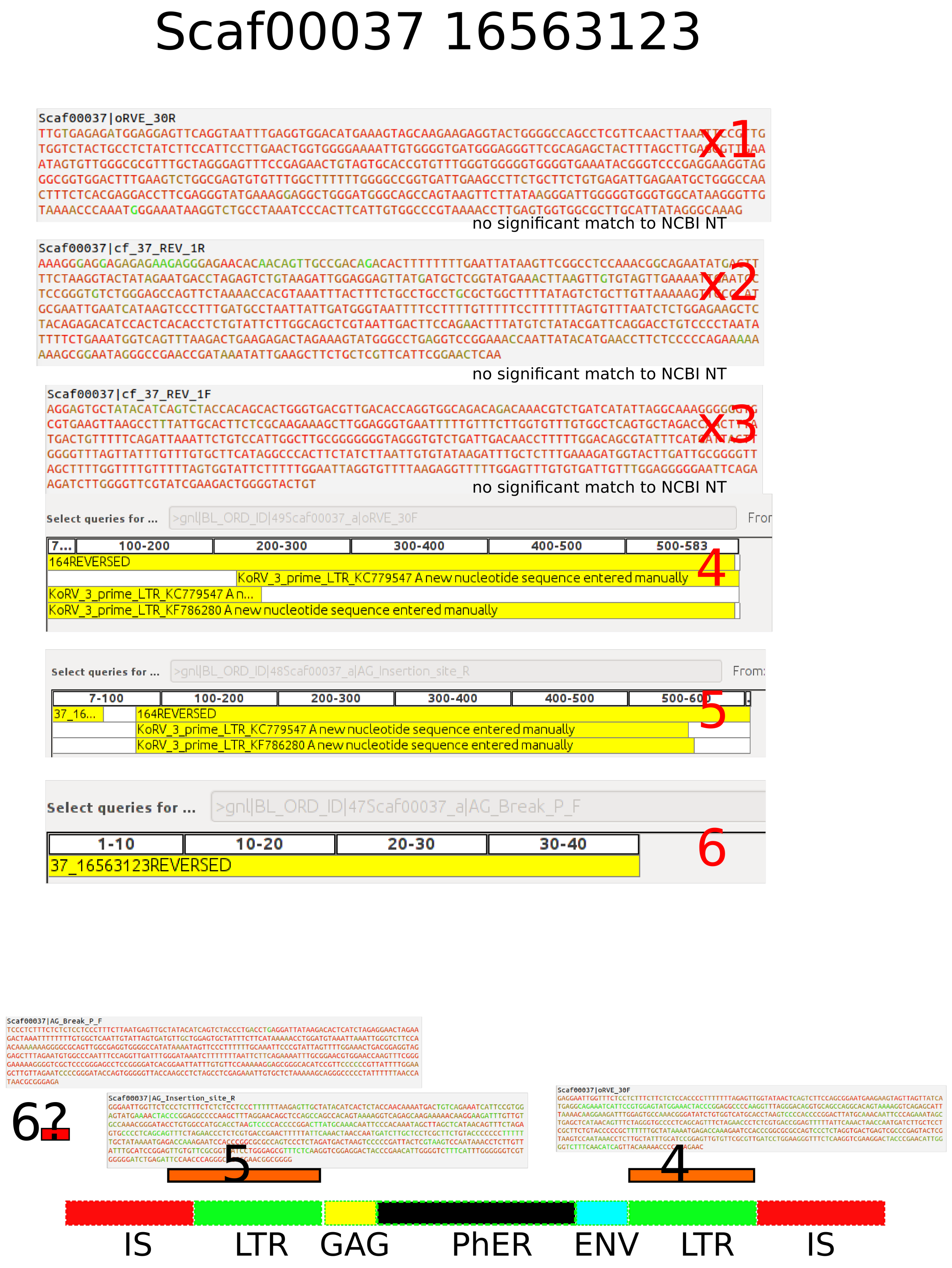
```
> Breakpoint_KoRVKoERV_Bi1
Length=101

 Score = 71.6 bits (78),  Expect = 8e-16
 Identities = 81/105 (77%), Gaps = 9/105 (9%)
 Strand=Plus/Minus

Query  289   GACCGAATAAAAAA-TCTTTCTTCTTTTTAGACCGTTTGGCTAGG---AACGATC-GATG  343
             |||| ||||||||| |||||||||||| |||||| |||  | |||   |||  ||   ||
Sbjct  101   GACCCAATAAAAAAATCTTTCTTCTTTCTAGACCCTTTCACCAGGTCCAACAGTCTCCTG  42

Query  344   GATGTTTGATAAACCTGTTGACTGCAGAAATAACAGAAGGATTAA  388
             ||     || || ||| |||||||||||||||| |||||||||||
Sbjct  41    GA----AGACAATCCTTTTGACTGCAGAAATAATAGAAGGATTAA  1
```

Figure A.4: Scaf37 16563123 PCR Products Sanger Sequencing



Scaf00037 16563123

```
Scaf00037|oRVE_30R
TTGTGAGAGATGGAGGAGTTCAGGTAATTTGAGGTGGACATGAAAGTAGCAAGAAGAGGTACTGGGGCCAGCCTCGTTCAACTTAAATTCCGTTG
TGGTCTACTGCCTCTATCTTCCATTCCTTGAACTGGTGGGGAAAATTGTGGGGTGATGGGAGGGTTCGCAGAGCTACTTTAGCTTGAGGGTTGAA
ATAGTGTTGGGCGCGTTTGCTAGGGAGTTTCCGAGAACTGTAGTGCACCGTGTTTGGGTGGGGGTGGGGTGAAATACGGGTCCCGAGGAAGGTAG
GGCGGTGGACTTTGAAGTCTGGCGAGTGTGTTTGGCTTTTTTGGGGCCGGTGATTGAAGCCTTCTGCTTCTGTGAGATTGAGAATGCTGGGCCAA
CTTTCTCACGAGGACCTTCGAGGGTATGAAAGGAGGCTGGGATGGGCAGCCAGTAAGTTCTTATAAGGGATTGGGGGTGGGTGGCATAAGGGTTG
TAAAACCCAAATGGGAAATAAGGTCTGCCTAAATCCCACTTCATTGTGGCCCGTAAAACCTTGAGTGGTGGCGCTTGCATTATAGGGCAAAG
```

no significant match to NCBI NT

```
Scaf00037|cf_37_REV_1R
AAAGGGAGGAGAGAGAAGAGGGAGAACACAACAGTTGCCGACAGACACTTTTTTTTGAATTATAAGTTCGGCCTCCAAACGGCAGAATATGAGTT
TTCTAAGGTACTATAGAATGACCTAGAGTCTGTAAGATTGGAGGAGTTATGATGCTCGGTATGAAACTTAAGTTGTGTAGTTGAAAATTGAATCC
TCCGGGTGTCTGGGAGCCAGTTCTAAAACCACGTAAATTTACTTTCTGCCTGCCTGCGCTGGCTTTTATAGTCTGCTTGTTAAAAAGTTCCGCAT
GCGAATTGAATCATAAGTCCCTTTGATGCCTAATTATTGATGGGTAATTTTCCTTTTGTTTTTCCTTTTTTAGTGTTTAATCTCTGGAGAAGCTC
TACAGAGACATCCACTCACACCTCTGTATTCTTGGCAGCTCGTAATTGACTTCCAGAACTTTATGTCTATACGATTCAGGACCTGTCCCCTAATA
TTTTCTGAAATGGTCAGTTTAAGACTGAAGAGACTAGAAAGTATGGGCCTGAGGTCCGGAAACCAATTATACATGAACCTTCTCCCCCAGAAAAA
AAAGCGGAATAGGGCCGAACCGATAAATATTGAAGCTTCTGCTCGTTCATTCGGAACTCAA
```

no significant match to NCBI NT

```
Scaf00037|cf_37_REV_1F
AGGAGTGCTATACATCAGTCTACCACAGCACTGGGTGACGTTGACACCAGGTGGCAGACAGACAAACGTCTGATCATATTAGGCAAAGGGGGGTG
CGTGAAGTTAAGCCTTTATTGCACTTCTCGCAAGAAAGCTTGGAGGGTGAATTTTTGTTTCTTGGTGTTTGTGGCTCAGTGCTAGACCAACTTTA
TGACTGTTTTTCAGATTAAATTCTGTCCATTGGCTTGCGGGGGGGTAGGGTGTCTGATTGACAACCTTTTTGGACAGCGTATTTCATGATTACTT
GGGGTTTAGTTATTTGTTTGTGCTTCATAGGCCCACTTCTATCTTAATTGTGTATAAGATTTGCTCTTTGAAAGATGGTACTTGATTGCGGGGTT
AGCTTTTGGTTTTGTTTTTAGTGGTATTCTTTTTGGAATTAGGTGTTTTAAGAGGTTTTTGGAGTTTGTGTGATTGTTTGGAGGGGGAATTCAGA
AGATCTTGGGGTTCGTATCGAAGACTGGGGTACTGT
```

no significant match to NCBI NT

Select queries for ... >gnl|BL_ORD_ID|49Scaf00037_a|oRVE_30F From

7...	100-200	200-300	300-400	400-500	500-583

164REVERSED

KoRV_3_prime_LTR_KC779547 A new nucleotide sequence entered manually

KoRV_3_prime_LTR_KC779547 A n...

KoRV_3_prime_LTR_KF786280 A new nucleotide sequence entered manually

Select queries for ... >gnl|BL_ORD_ID|48Scaf00037_a|AG_Insertion_site_R From:

7-100	100-200	200-300	300-400	400-500	500-60

37_16...

164REVERSED

KoRV_3_prime_LTR_KC779547 A new nucleotide sequence entered manually

KoRV_3_prime_LTR_KF786280 A new nucleotide sequence entered manually

Select queries for ... >gnl|BL_ORD_ID|47Scaf00037_a|AG_Break_P_F

1-10	10-20	20-30	30-40

37_16563123REVERSED

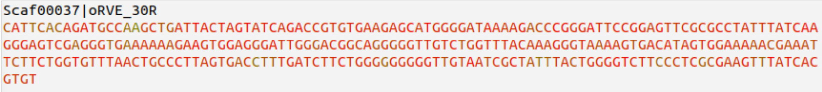
```
Scaf00037|AG_Break_P_F
TCCCTCTTTCTCTCTCCTCCCTTTCTTAATGAGTTGCTATACATCAGTCTACCCTGACCTGAGGATTATAAGACACTCATCTAGAGGAACTAGAA
GACTAAATTTTTTTTGTGGCTCAATTGTATTAGTGATGTTGCTGGAGTGCTATTTCTTCATAAAAACCTGGATGTAAATTAAATTGGGTCTTCCA
ACAAAAAAAGGGGCGCAGTTGGCGAGGTGGGGCCATATAAAATAGTTCCCTTTTTTGCAAATTCCCGTATTAGTTTTGGAAACTGACGGAGGTAG
GAGCTTTAGAATGTGGCCCAATTTCCAGGTTGATTTGGGATAAATCTTTTTTTAATTCTTCAGAAAATTTGCGGAACGTGGAACCAAGTTTCGGG
GAAAAAGGGGTCGCTCCCGGGAGCCTCCGGGGATCACGGAATTATTTGTGTTCCAAAAAGGAGCGGGCACATCCGTTCCCCCCGTTATTTTGGAA
GCTTGTTAGAATCCCCGGGATACCAGTGGGGGTTACCAAGCCTCTAGCCTCGAGAAATTGTGCTCTAAAAAGCAGGGCCCCCTATTTTTTAACCA
TAACGCGGGAGA
```

```
Scaf00037|AG_Insertion_site_R
GGGAATTGGTTCTCCCTCTTTCTCTCTCCTCCCTTTTTTAAGAGTTGCTATACATCACTCTACCAACAAAATGACTGTCAGAAATCATTCCGTGG
AGTATGAAAACTACCCGGAGGCCCCAAGCTTTAGGAACAGCTCCAGCCAGCCACAGTAAAAGGTCAGAGCAAGAAAAACAAGGAAGATTTGTTGT
GCCAAACGGGATACCTGTGGCCATGCACCTAAGTCCCCACCCCGGACTTATGCAAACAATTCCCACAAATAGCTTAGCTCATAACAGTTTCTAGA
GTGCCCCTCAGCAGTTTCTAGAACCCTCTCGTGACCGAACTTTTTATTCAAACTAACCAATGATCTTGCTCCTCGCTTCTGTACCCCCCCTTTTT
TGCTATAAAATGAGACCAAAGAATCCACCCGGCGCGCCAGTCCCTCTAGATGACTAAGTCCCCCGATTACTCGTAAGTCCAATAAACCTCTTGTT
ATTTGCATCCGGAGTTGTGTTCGCGGTTATCCTGGGAGCGTTTCTCAAGGTCGGAGGACTACCCGAACATTGGGGTCTTTCATTTGGGGGGTCGT
GGGGGATCTGAGATTCCAACCCAGGGCGGCGAACGGCGGGG
```

```
Scaf00037|oRVE_30F
GAGGAATTGGTTTCTCCTCTTTCTTCTCTCCACCCCTTTTTTTAGAGTTGGTATAACTCAGTCTTCCAGCGGAATGAAGAAGTAGTTAGTTATCA
TGAGGCAGAAATCATTCCGTGGAGTATGGAAACTACCCGGAGGCCCCAAGGTTTAGGGACAGGTGCAGCCAGGCACAGTAAAAGGTCAGAGCATT
TAAAACAAGGAAGATTTGGAGTGCCAAACGGGATATCTGTGGTCATGCACCTAAGTCCCCACCCCGGACTTATGCAAACAATTCCCAGAAATAGC
TGAGCTCATAACAGTTTCTAGGGTGCCCCTCAGCAGTTTCTAGAACCCTCTCGTGACCGGAGTTTTTATTCAAACTAACCAATGATCTTGCTCCT
CGCTTCTGTACCCCCGCTTTTTTGCTATAAAATGAGACCAAAGAATCCACCCGGCGCGCCAGTCCCTCTAGGTGACTGAGTCGCCCGAGTACTCG
TAAGTCCAATAAACCTCTTGCTATTTGCATCCGGAGTTGTGTTCGCGTTGATCCTGGAAGGGTTTCTCAAGGTCGAAGGACTACCCGAACATTGG
GGTCTTTCAACATCAGTTACAAAAACCCCTAGAAC
```

Figure A.5: Scaf37 17582840 PCR Products Sanger Sequencing

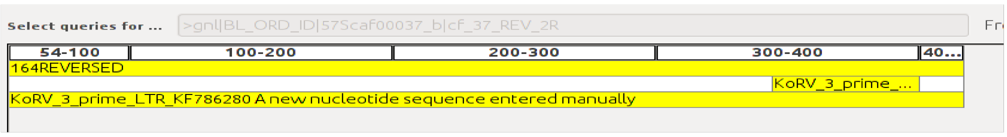
Scaf00037 17582840



Scaf00037|oRVE_30R
CATTCACAGATGCCAAGCTGATTACTAGTATCAGACCGTGTGAAGAGCATGGGGATAAAAGACCCGGGATTCCGGAGTTCGCGCCTATTTATCAA
GGGAGTCGAGGGTGAAAAAAGAAGTGGAGGGATTGGGACGGCAGGGGGTTGTCTGGTTTACAAAGGGTAAAAGTGACATAGTGGAAAAACGAAAT
TCTTCTGGTGTTTAACTGCCCTTAGTGACCTTTGATCTTCTGGGGGGGGGTTGTAATCGCTATTTACTGGGGTCTTCCCTCGCGAAGTTTATCAC
GTGT

x1

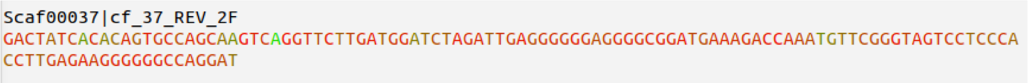
no significant match to NCBI NT



Select queries for ... >gnl|BL_ORD_ID|57Scaf00037_b|cf_37_REV_2R Fr

54-100 | 100-200 | 200-300 | 300-400 | 40...
164REVERSED
KoRV_3_prime_...
KoRV_3_prime_LTR_KF786280 A new nucleotide sequence entered manually

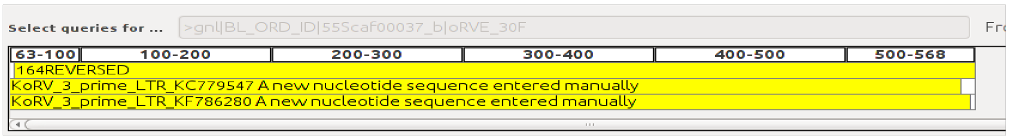
2



Scaf00037|cf_37_REV_2F
GACTATCACACAGTGCCAGCAAGTCAGGTTCTTGATGGATCTAGATTGAGGGGGGAGGGGCGGATGAAAGACCAAATGTTCGGGTAGTCCTCCCA
CCTTGAGAAGGGGGGCCAGGAT

x3

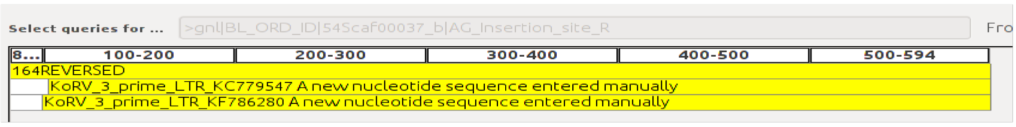
no significant match to NCBI NT



Select queries for ... >gnl|BL_ORD_ID|55Scaf00037_b|oRVE_30F Fr

63-100 | 100-200 | 200-300 | 300-400 | 400-500 | 500-568
164REVERSED
KoRV_3_prime_LTR_KC779547 A new nucleotide sequence entered manually
KoRV_3_prime_LTR_KF786280 A new nucleotide sequence entered manually

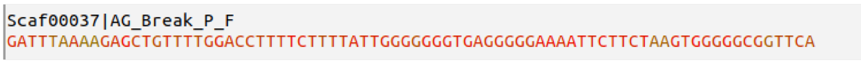
4



Select queries for ... >gnl|BL_ORD_ID|54Scaf00037_b|AG_Insertion_site_R Fro

8... | 100-200 | 200-300 | 300-400 | 400-500 | 500-594
164REVERSED
KoRV_3_prime_LTR_KC779547 A new nucleotide sequence entered manually
KoRV_3_prime_LTR_KF786280 A new nucleotide sequence entered manually

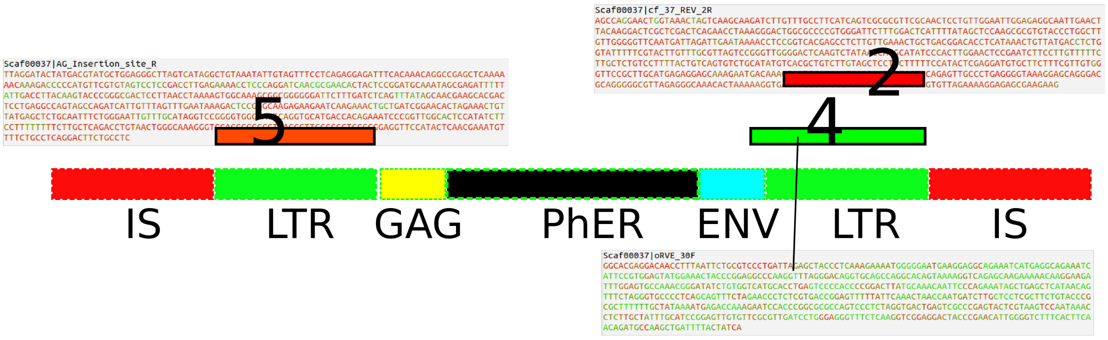
5



Scaf00037|AG_Break_P_F
GATTTAAAAGAGCTGTTTTGGACCTTTTCTTTTATTGGGGGGGTGAGGGGGAAAATTCTTCTAAGTGGGGGCGGTTCA

x6

no significant match to NCBI NT

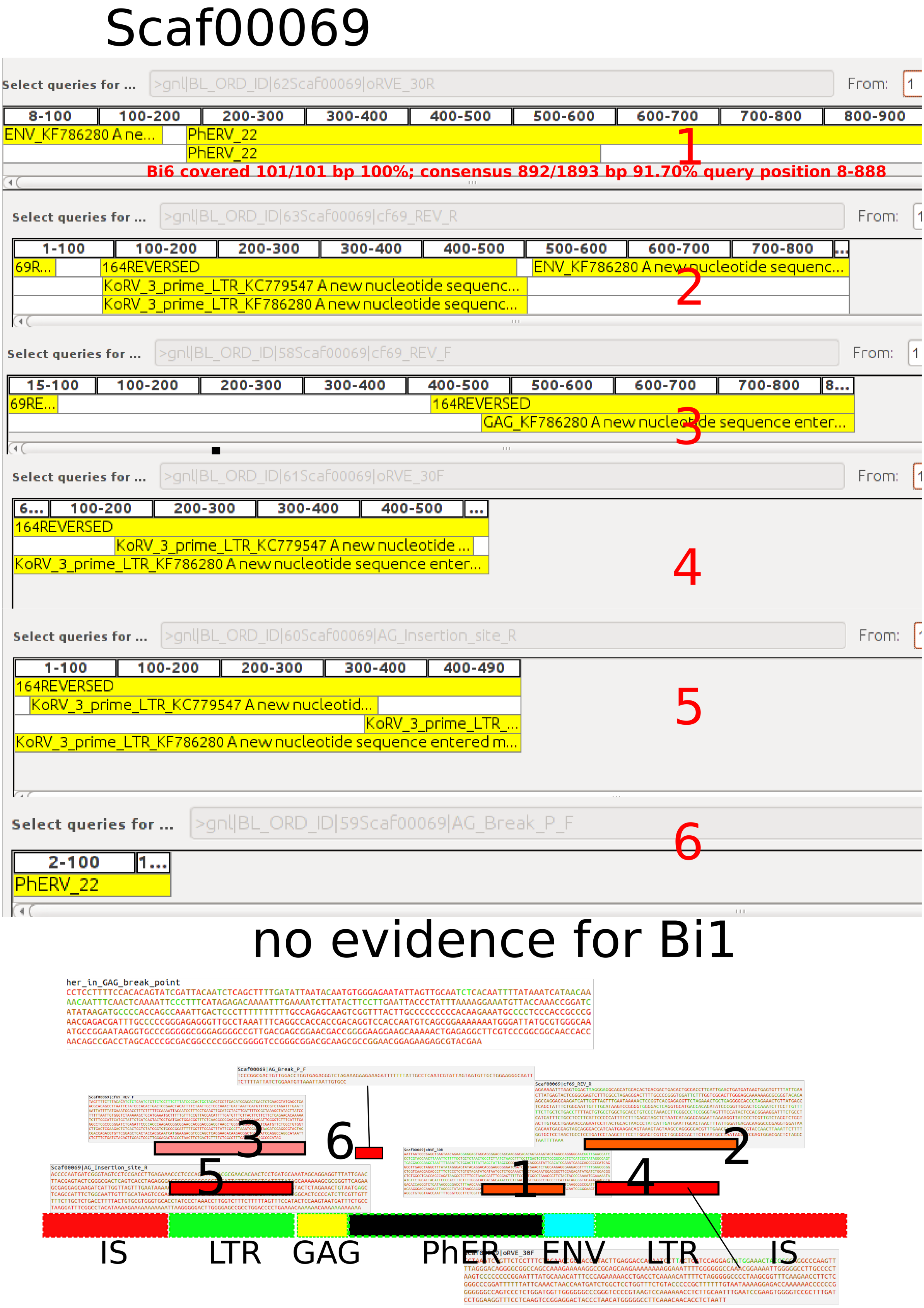


Scaf00037|cf_37_REV_2R
AGCCAGGAACTGGTAAACTAGTCAAGCAAGATCTTGTTTGCCTTCATCAGTCGCGCGTTCGCAACTCCTGTTGGAATTGGAGAGGCAATTGAACT
TACAAGGACTCGCTCGACTCAGAACCTAAAGGGACTGGCGCCCCGTGGGATTCTTTGGACTCATTTTATAGCTCCAAGCGCGTGTACCCTGGCTT
GTTGGGGGTTCAATGATTAGATTGAATAAAACCTCCGGTCACGAGCCTCTTGTTGAAACTGCTGACGGACACCTCATAAACTGTTATGACCTCTG
GTATTTTTCGTACTTGTTTGCGTTAGTCCGGGTTGGGGACTCAAGTCTATAA[illegible]GCATATCCCACTTGGAACTCCGAATCTTCCTTGTTTTTC
TTGCTCTGTCCTTTTACTGTCAGTGTCTGCATATGTCACGCTGTCTTGTAGCTCCT[illegible]CTTTTTCCATACTCGAGGATGTGCTTCTTTCGTTGTGG
GTTCCGCTTGCATGAGAGGAGCAAAGAATGACAAA[illegible]CAGAGTTGCCCTGAGGGGTAAAGGAGCAGGGAC
GCAGGGGGCGTTAGAGGGCAAACACTAAAAAGGTG[illegible]GTGTTAGAAAAGGAGAGCGAAGAAG

Scaf00037|AG_Insertion_site_R
TTAGGATACTATGACGTATGCTGGAGGGCTTAGTCATAGGCTGTAAATATTGTAGTTTCCTCAGAGGAGATTTCACAAACAGGCCGAGCTCAAAA
AACAAAGACCCCCATGTTCGTGTAGTCCTCCGACCTTGAGAAAACCTCCCAGGATCAACGCGAACACTACTCCGGATGCAAATAGCGAGATTTTT
ATTGACCTTACAAGTACCCGGGCGACTCCTTAACCTAAAAGTGGCAAAGCGGCGGGGGGATTCTTTGATCTCAGTTTATAGCAACGAAGCACGAC
TCCTGAGGCCAGTAGCCAGATCATTGTTTAGTTTGAATAAAGACTCCC[illegible]GCAAGAGAAGAATCAAGAAACTGCTGATCGGAACACTAGAAACTGT
TATGAGCTCTGCAATTTCTGGGAATTGTTTGCATAGGTCCGGGGTGGC[illegible]CAGGTGCATGACCACAGAAATCCCGGTTGGCACTCCATATCTT
CCTTTTTTTTCTTGCTCAGACCTGTAACTGGGCAAAGGT[illegible]GGAGGTTCCATACTCAACGAAATGT
TTTCTGCCTCAGGACTTCTGCCTC

Scaf00037|oRVE_30F
GGCACGAGGACAACCTTTAATTCTGCGTCCCTGATTAGAGCTACCCTCAAAGAAAATGGGGAATGAAGGAGGCAGAAATCATGAGGCAGAAATC
ATTCCGTGGAGTATGGAAACTACCCGGAGGCCCAAGGTTTAGGGACAGGTGCAGCCAGGCACAGTAAAAGGTCAGAGCAAGAAAAACAAGGAAGA
TTTGGAGTGCCAAACGGGATATCTGTGGTCATGCACCTGAGTCCCCACCCCGGACTTATGCAAACAATTCCCAGAAATAGCTGAGCTCATAACAG
TTTCTAGGGTGCCCCTCAGCAGTTTCTAGAACCCTCTCGTGACCGGAGTTTTTATTCAAACTAACCAATGATCTTGCTCCTCGCTTCTGTACCCG
CGCTTTTTTGCTATAAAATGAGACCAAAGAATCCACCCGGCGCGCCAGTCCCTCTAGGTGACTGAGTCGCCCGAGTACTCGTAAGTCCAATAAAC
CTCTTGCTATTTGCATCCGGAGTTGTGTTCGCGTTGATCCTGGGAGGGTTTCTCAAGGTCGGAGGACTACCCGAACATTGGGGTCTTTCACTTCA
ACAGATGCCAAGCTGATTTTACTATCA

Figure A.6: Scaf69 PCR Products Sanger Sequencing



Scaf00069

Select queries for ... >gnl|BL_ORD_ID|62Scaf00069|oRVE_30R From: 1

8-100
100-200
200-300
300-400
400-500
500-600
600-700
700-800
800-900
ENV_KF786280 A ne...
PhERV_22
PhERV_22
1

Bi6 covered 101/101 bp 100%; consensus 892/1893 bp 91.70% query position 8-888

Select queries for ... >gnl|BL_ORD_ID|63Scaf00069|cf69_REV_R From: 1

1-100
100-200
200-300
300-400
400-500
500-600
600-700
700-800
69R...
164REVERSED
ENV_KF786280 A new nucleotide sequenc...
KoRV_3_prime_LTR_KC779547 A new nucleotide sequenc...
KoRV_3_prime_LTR_KF786280 A new nucleotide sequenc...
2

Select queries for ... >gnl|BL_ORD_ID|58Scaf00069|cf69_REV_F From: 1

15-100
100-200
200-300
300-400
400-500
500-600
600-700
700-800
8...
69RE...
164REVERSED
GAG_KF786280 A new nucleotide sequence enter...
3

Select queries for ... >gnl|BL_ORD_ID|61Scaf00069|oRVE_30F From: 1

6...
100-200
200-300
300-400
400-500
...
164REVERSED
KoRV_3_prime_LTR_KC779547 A new nucleotide ...
KoRV_3_prime_LTR_KF786280 A new nucleotide sequence enter...
4

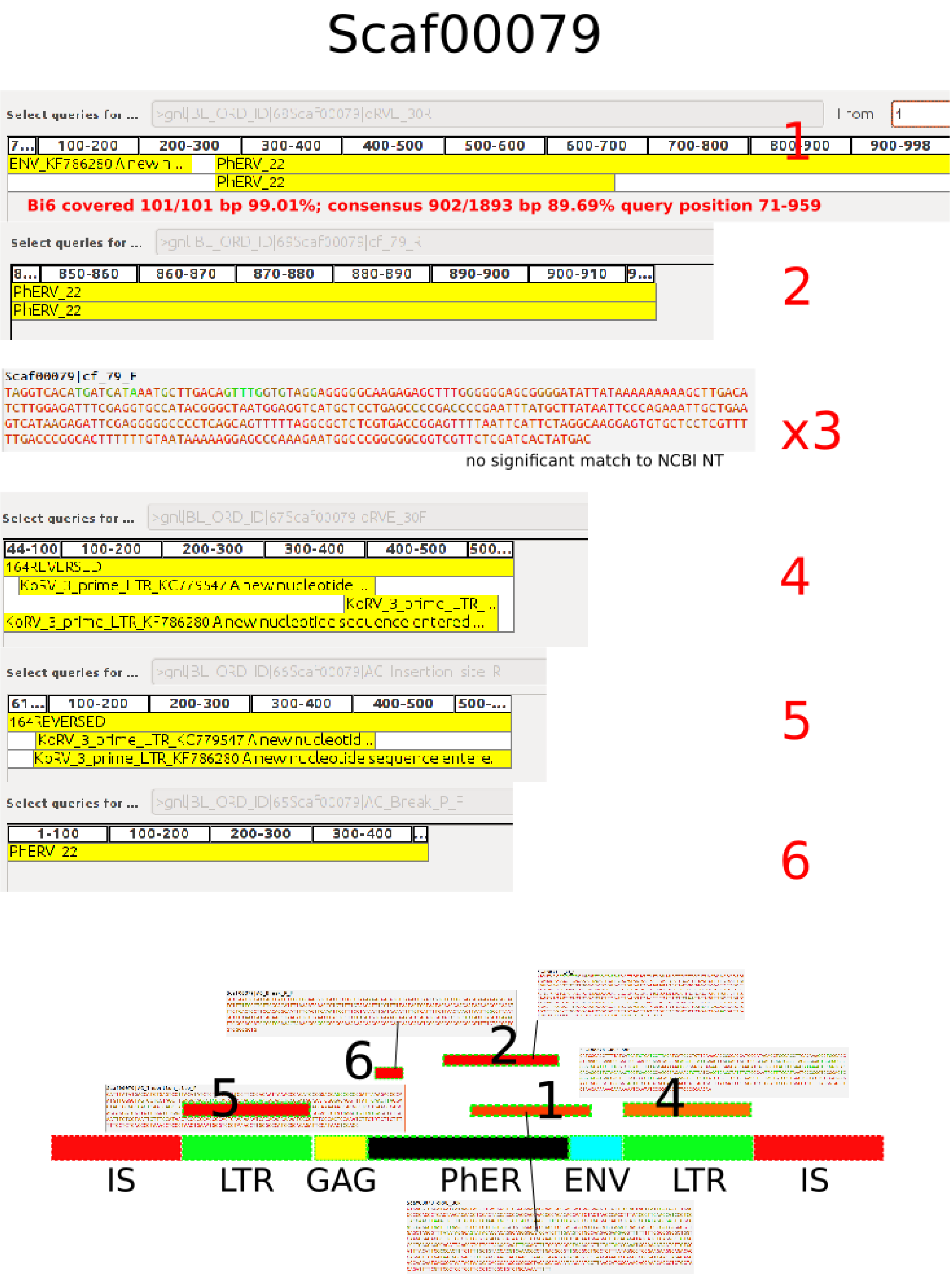
Select queries for ... >gnl|BL_ORD_ID|60Scaf00069|AG_Insertion_site_R From: 1

1-100
100-200
200-300
300-400
400-490
164REVERSED
KoRV_3_prime_LTR_KC779547 A new nucleotid...
KoRV_3_prime_LTR_...
KoRV_3_prime_LTR_KF786280 A new nucleotide sequence entered m...
5

Select queries for ... >gnl|BL_ORD_ID|59Scaf00069|AG_Break_P_F

no evidence for Bi1

Figure A.7: Scaf79 PCR Products Sanger Sequencing



Scaf00079

Select queries for ... >gnl|BL_ORD_ID|69Scaf00079|cRVL_30R From 1

7...	100-200	200-300	300-400	400-500	500-600	600-700	700-800	800-900	900-998

ENV_KF786280 A new n ..
PhERV_22
PhERV_22

Bi6 covered 101/101 bp 99.01%; consensus 902/1893 bp 89.69% query position 71-959

Select queries for ... >gnl|BL_ORD_ID|69Scaf00079|cf_79_R

8...	850-860	860-870	870-880	880-890	890-900	900-910	9...

PhERV_22
PhERV_22

```
Scaf00079|cf_79_F
TAGGTCACATGATCATAAATGCTTGACAGTTTGGTGTAGGAGGGGGCAAGAGAGCTTTGGGGGGAGCGGGGATATTATAAAAAAAAAGCTTGACA
TCTTGGAGATTTCGAGGTGCCATACGGGCTAATGGAGGTCATGCTCCTGAGCCCCGACCCCGAATTTATGCTTATAATTCCCAGAAATTGCTGAA
GTCATAAGAGATTCGAGGGGGCCCCTCAGCAGTTTTTAGGCGCTCTCGTGACCGGAGTTTTAATTCATTCTAGGCAAGGAGTGTGCTCCTCGTTT
TTGACCCGGCACTTTTTTGTAATAAAAAGGAGCCCAAAGAATGGCCCGGCGGCGGTCGTTCTCGATCACTATGAC
```

no significant match to NCBI NT

Select queries for ... >gnl|BL_ORD_ID|67Scaf00079 cRVE_30F

44-100	100-200	200-300	300-400	400-500	500...

164REVERSED
KoRV_5_prime_LTR_KC779547 A new nucleotide ..
KoRV_3_prime_LTR_ ..
KoRV_3_prime_LTR_KF786280 A new nucleotide sequence entered ...

Select queries for ... >gnl|BL_ORD_ID|65Scaf00079|AC Insertion site R

61 ...	100-200	200-300	300-400	400-500	500-...

164REVERSED
KoRV_3_prime_LTR_KC779547 A new nucleotid ..
KoRV_3_prime_LTR_KF786280 A new nucleotide sequence entere.

Select queries for ... >gnl|BL_ORD_ID|65Scaf00079|AC_Break_P_F

1-100	100-200	200-300	300-400	..

PHERV 22

Figure A.8: Scaf83 PCR Products Sanger Sequencing

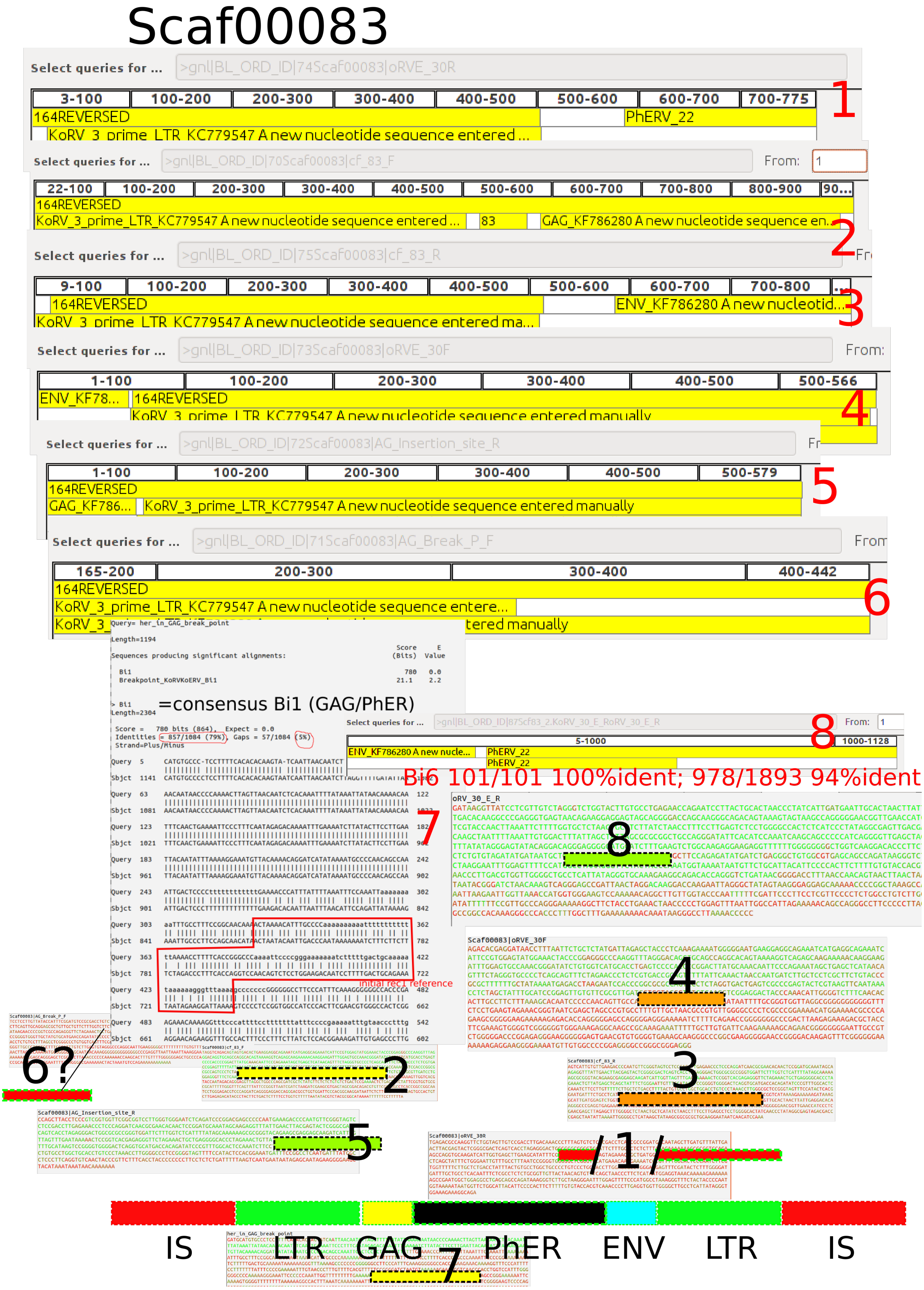
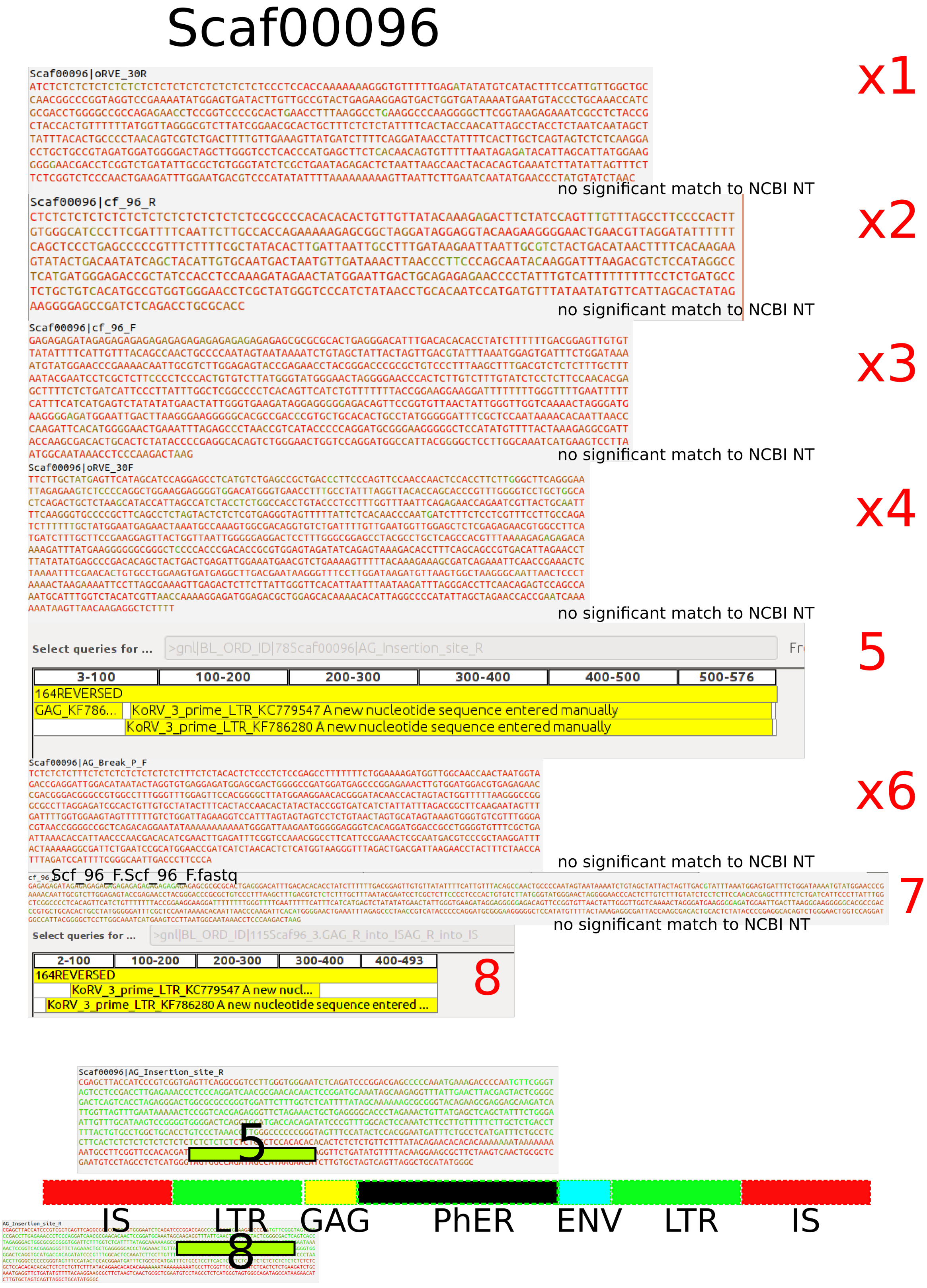

Scaf00083
Select queries for ...
>gnl|BL_ORD_ID|74Scaf00083|oRVE_30R
3-100
100-200
200-300
300-400
400-500
500-600
600-700
700-775
164REVERSED
PhERV_22
KoRV_3_prime_LTR_KC779547 A new nucleotide sequence entered ...
1
>gnl|BL_ORD_ID|70Scaf00083|cf_83_F
From: 1
22-100
100-200
200-300
300-400
400-500
500-600
600-700
700-800
800-900
90...
164REVERSED
KoRV_3_prime_LTR_KC779547 A new nucleotide sequence entered ...
83
GAG_KF786280 A new nucleotide sequence en...
2
>gnl|BL_ORD_ID|75Scaf00083|cf_83_R
9-100
100-200
200-300
300-400
400-500
500-600
600-700
700-800
164REVERSED
ENV_KF786280 A new nucleotid...
KoRV_3_prime_LTR_KC779547 A new nucleotide sequence entered ma...
3
>gnl|BL_ORD_ID|73Scaf00083|oRVE_30F
From:
1-100
100-200
200-300
300-400
400-500
500-566
ENV_KF78...
164REVERSED
KoRV_3_prime_LTR_KC779547 A new nucleotide sequence entered manually
4
>gnl|BL_ORD_ID|72Scaf00083|AG_Insertion_site_R
1-100
100-200
200-300
300-400
400-500
500-579
164REVERSED
GAG_KF786...
KoRV_3_prime_LTR_KC779547 A new nucleotide sequence entered manually
5
>gnl|BL_ORD_ID|71Scaf00083|AG_Break_P_F
165-200
200-300
300-400
400-442
164REVERSED
KoRV_3_prime_LTR_KC779547 A new nucleotide sequence entere...
KoRV_3
tered manually
6
Query= her_in_GAG_break_point
Length=1194
Sequences producing significant alignments:
Score (Bits)
E Value
Bi1
780
0.0
Breakpoint_KoRVKoERV_Bi1
21.1
2.2
> Bi1
Length=2304
=consensus Bi1 (GAG/PhER)
Score = 780 bits (864), Expect = 0.0
Identities = 857/1084 (79%), Gaps = 57/1084 (5%)
Strand=Plus/Minus
initial rec1 reference
7
>gnl|BL_ORD_ID|87Scf83_2.KoRV_30_E_RoRV_30_E_R
From: 1
5-1000
1000-1128
ENV_KF786280 A new nucle...
PhERV_22
PhERV_22
8
Bi6 101/101 100%ident; 978/1893 94%ident
oRV_30_E_R
Scaf00083|oRVE_30F
Scaf00083|cf_83_R
Scaf00083|AG_Break_P_F
Scaf00083|AG_Insertion_site_R
Scaf00083|cf_83_F
Scaf00083|oRVE_30R
her_in_GAG_break_point
8
4
3
2
6?
5
/1/
7
IS
LTR
GAG
PhER
ENV
LTR
IS

Figure A.9: Scaf96 PCR Products Sanger Sequencing



```
Scaf00096|oRVE_30R
ATCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCCCTCCACCAAAAAAAGGGTGTTTTTGAGATATATGTCATACTTTCCATTGTTGGCTGC
CAACGGCCCGGTAGGTCCGAAAATATGGAGTGATACTTGTTGCCGTACTGAGAAGGAGTGACTGGTGATAAAATGAATGTACCCTGCAAACCATC
GCGACCTGGGGCCGCCAGAGAACCTCCGGTCCCCGCACTGAACCTTTAAGGCCTGAAGGCCCAAGGGGCTTCGGTAAGAGAAATCGCCTCTACCG
CTACCACTGTTTTTTATGGTTAGGGCGTCTTATCGGAACGCACTGCTTTCTCTCTATTTTCACTACCAACATTAGCCTACCTCTAATCAATAGCT
TATTTACACTGCCCCTAACAGTCGTCTGACTTTTGTTGAAAGTTATGATCTTTTCAGGATAACCTATTTTCACTTGCTCAGTAGTCTCTCAAGGA
CCTGCTGCCGTAGATGGATGGGGACTAGCTTGGGTCCTCACCCATGAGCTTCTCACAACAGTGTTTTTAATAGAGATACATTAGCATTATGGAAG
GGGGAACGACCTCGGTCTGATATTGCGCTGTGGGTATCTCGCTGAATAGAGACTCTAATTAAGCAACTACACAGTGAAATCTTATATTAGTTTCT
TCTCGGTCTCCCAACTGAAGATTTGGAATGACGTCCCATATATTTTAAAAAAAAAGTTAATTCTTGAATCAATATGAACCCTATGTATCTAAC
```

no significant match to NCBI NT

```
Scaf00096|cf_96_R
CTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCCGCCCCACACACACTGTTGTTATACAAAGAGACTTCTATCCAGTTTGTTTAGCCTTCCCCACTT
GTGGGCATCCCTTCGATTTTCAATTCTTGCCACCAGAAAAAGAGCGGCTAGGATAGGAGGTACAAGAAGGGGAACTGAACGTTAGGATATTTTTT
CAGCTCCCTGAGCCCCCGTTTCTTTTCGCTATACACTTGATTAATTGCCTTTGATAAGAATTAATTGCGTCTACTGACATAACTTTTCACAAGAA
GTATACTGACAATATCAGCTACATTGTGCAATGACTAATGTTGATAAACTTAACCCTTCCCAGCAATACAAGGATTTAAGACGTCTCCATAGGCC
TCATGATGGGAGACCGCTATCCACCTCCAAAGATAGAACTATGGAATTGACTGCAGAGAGAACCCCTATTTGTCATTTTTTTTTCCTCTGATGCC
TCTGCTGTCACATGCCGTGGTGGGAACCTCGCTATGGGTCCCATCTATAACCTGCACAATCCATGATGTTTATAATATGTTCATTAGCACTATAG
AAGGGGAGCCGATCTCAGACCTGCGCACC
```

no significant match to NCBI NT

```
Scaf00096|cf_96_F
GAGAGAGATAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGCGCGCACTGAGGGACATTTGACACACACCTATCTTTTTTGACGGAGTTGTGT
TATATTTTCATTGTTTACAGCCAACTGCCCCAATAGTAATAAAATCTGTAGCTATTACTAGTTGACGTATTTAAATGGAGTGATTTCTGGATAAA
ATGTATGGAACCCGAAAACAATTGCGTCTTGGAGAGTACCGAGAACCTACGGGACCCGCGCTGTCCCTTTAAGCTTTGACGTCTCTCTTTGCTTT
AATACGAATCCTCGCTCTTCCCCTCCCACTGTGTCTTATGGGTATGGGAACTAGGGGAACCCACTCTTGTCTTTGTATCTCCTCTTCCAACACGA
GCTTTTCTCTGATCATTCCCTTATTTGGCTCGGCCCCTCACAGTTCATCTGTTTTTTTACCGGAAGGAAGGATTTTTTTTGGGTTTTGAATTTTT
CATTTCATCATGAGTCTATATATGAACTATTGGGTGAAGATAGGAGGGGGAGACAGTTCCGGTGTTAACTATTGGGTTGGTCAAAACTAGGGATG
AAGGGGAGATGGAATTGACTTAAGGAAGGGGGCACGCCGACCCGTGCTGCACACTGCCTATGGGGGATTTCGCTCCAATAAAACACAATTAACC
CAAGATTCACATGGGGAACTGAAATTTAGAGCCCTAACCGTCATACCCCCAGGATGCGGGAAGGGGGCTCCATATGTTTTACTAAAGAGGCGATT
ACCAAGCGACACTGCACTCTATACCCCGAGGCACAGTCTGGGAACTGGTCCAGGATGGCCATTACGGGGCTCCTTGGCAAATCATGAAGTCCTTA
ATGGCAATAAACCTCCCAAGACTAAG
```

no significant match to NCBI NT

```
Scaf00096|oRVE_30F
TTCTTGCTATGAGTTCATAGCATCCAGGAGCCTCATGTCTGAGCCGCTGACCCTTCCCAGTTCCAACCAACTCCACCTTCTTGGGCTTCAGGGAA
TTAGAGAAGTCTCCCCAGGCTGGAAGGAGGGGTGGACATGGGTGAACCTTTGCCTATTTAGGTTACACCAGCACCCGTTTGGGGTCCTGCTGGCA
CTCAGACTGCTCTAAGCATACCATTAGCCATCTACCTCTGGCCACCTGTACCCTCCTTTGGTTTAATTCAGAGAACCAGAATCGTTACTGCAATT
TTCAAGGGTGCCCCGCTTCAGCCTCTAGTACTCTCTCGTGAGGGTAGTTTTTATTCTCACAACCCAATGATCTTTCTCCTCGTTTCCTTGCCAGA
TCTTTTTTGCTATGGAATGAGAACTAAATGCCAAAGTGGCGACAGGTGTCTGATTTTGTTGAATGGTTGGAGCTCTCGAGAGAACGTGGCCTTCA
TGATCTTTGCTTCCGAAGGAGTTACTGGTTAATTGGGGGAGGACTCCTTTGGGCGGAGCCTACGCCTGCTCAGCCACGTTTAAAAGAGAGAGACA
AAAGATTTATGAAGGGGGGCGGGCTCCCCACCCGACACCGCGTGGAGTAGATATCAGAGTAAAGACACCTTTCAGCAGCCGTGACATTAGAACCT
TTATATATGAGCCCGACACAGCTACTGACTGAGATTGGAAATGAACGTCTGAAAAGTTTTTACAAAGAAAGCGATCAGAAATTCAACCGAAACTC
TAAAATTTCGAACACTGTGCCTGGAAGTGATGAGGCTTGACGAATAAGGGTTTCCTTGGATAAGATGTTAAGTGGCTAAGGGCAATTAACTCCCT
AAAACTAAGAAAATTCCTTAGCGAAAGTTGAGACTCTTCTTATTGGGTTCACATTAATTTAATAAGATTTAGGGACCTTCAACAGAGTCCAGCCA
AATGCATTTGGTCTACATCGTTAACCAAAAGGAGATGGAGACGCTGGAGCACAAAACACATTAGGCCCCATATTAGCTAGAACCACCGAATCAAA
AAATAAGTTAACAAGAGGCTCTTTT
```

no significant match to NCBI NT

Select queries for ... >gnl|BL_ORD_ID|78Scaf00096|AG_Insertion_site_R Fr

3-100	100-200	200-300	300-400	400-500	500-576
164REVERSED					
GAG_KF786...	KoRV_3_prime_LTR_KC779547 A new nucleotide sequence entered manually				
	KoRV_3_prime_LTR_KF786280 A new nucleotide sequence entered manually				

```
Scaf00096|AG_Break_P_F
TCTCTCTCTTTCTCTCTCTCTCTCTCTCTTTCTCTACACTCTCCCTCTCCGAGCCTTTTTTTCTGGAAAAGATGGTTGGCAACCAACTAATGGTA
GACCGAGGATTGGACATAATACTAGGTGTGAGGAGATGGAGCGACTGGGGCCGATGGATGAGCCCGGAGAAACTTGTGGATGGACGTGAGAGAAC
CGACGGGACGGGCCGTGGCCTTTGGGTTTGGAGTTCCACGGGGCTTATGGAAGGAACACGGGATACAACCACTAGTACTGGTTTTTAAGGGCCGG
GCGCCTTAGGAGATCGCACTGTTGTGCTATACTTTCACTACCAACACTATACTACCGGTGATCATCTATTATTTAGACGGCTTCAAGAATAGTTT
GATTTTGGTGGAAGTAGTTTTTTGTCTGGATTAGAAGGTCCATTTAGTAGTAGTCCTCTGTAACTAGTGCATAGTAAAGTGGGTGTCGTTTGGGA
CGTAACCGGGGCCGCTCAGACAGGAATATAAAAAAAAAAATGGGATTAAGAATGGGGGAGGGTCACAGGATGGACCGCCTGGGGTGTTTCGCTGA
ATTAAACACCATTAACCCAACGACACATCGAACTTGAGATTTCGGTCCAAACGGCCTTCATTCCGAAACTCGCAATGACGTCCCGCTAAGGATTT
ACTAAAAAGGCGATTCTGAATCCGCATGGAACCGATCATCTAACACTCTCATGGTAAGGGTTTAGACTGACGATTAAGAACCTACTTTCTAACCA
TTTAGATCCATTTTCGGGCAATTGACCCTTCCCA
```

no significant match to NCBI NT

Scf_96_F.Scf_96_F.fastq

```
cf_96_
GAGAGAGATAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGCGCGCACTGAGGGACATTTGACACACACCTATCTTTTTTGACGGAGTTGTGTTATATTTTCATTGTTTACAGCCAACTGCCCCAATAGTAATAAAATCTGTAGCTATTACTAGTTGACGTATTTAAATGGAGTGATTTCTGGATAAAATGTATGGAACCCG
AAAACAATTGCGTCTTGGAGAGTACCGAGAACCTACGGGACCCGCGCTGTCCCTTTAAGCTTTGACGTCTCTCTTTGCTTTAATACGAATCCTCGCTCTTCCCCTCCCACTGTGTCTTATGGGTATGGGAACTAGGGGAACCCACTCTTGTCTTTGTATCTCCTCTTCCAACACGAGCTTTTCTCTGATCATTCCCTTATTTGG
CTCGGCCCCTCACAGTTCATCTGTTTTTTTACCGGAAGGAAGGATTTTTTTTGGGTTTTGAATTTTTCATTTCATCATGAGTCTATATATGAACTATTGGGTGAAGATAGGAGGGGGAGACAGTTCCGGTGTTAACTATTGGGTTGGTCAAAACTAGGGATGAAGGGGAGATGGAATTGACTTAAGGAAGGGGGCACGCCGAC
CCGTGCTGCACACTGCCTATGGGGGATTTCGCTCCAATAAAACACAATTAACCCAAGATTCACATGGGGAACTGAAATTTAGAGCCCTAACCGTCATACCCCCAGGATGCGGGAAGGGGGCTCCATATGTTTTACTAAAGAGGCGATTACCAAGCGACACTGCACTCTATACCCCGAGGCACAGTCTGGGAACTGGTCCAGGAT
GGCCATTACGGGGCTCCTTGGCAAATCATGAAGTCCTTAATGGCAATAAACCTCCCAAGACTAAG
```

no significant match to NCBI NT

Select queries for ... >gnl|BL_ORD_ID|115Scaf96_3.GAG_R_into_ISAG_R_into_IS

2-100	100-200	200-300	300-400	400-493
164REVERSED				
	KoRV_3_prime_LTR_KC779547 A new nucl...			
KoRV_3_prime_LTR_KF786280 A new nucleotide sequence entered ...				

```
Scaf00096|AG_Insertion_site_R
CGAGCTTACCATCCCGTCGGTGAGTTCAGGCGGTCCTTGGGTGGGAATCTCAGATCCCGGACGAGCCCCCAAATGAAAGACCCCAATGTTCGGGT
AGTCCTCCGACCTTGAGAAACCCTCCCAGGATCAACGCGAACACAACTCCGGATGCAAATAGCAAGAGGTTTATTGAACTTACGAGTACTCGGGC
GACTCAGTCACCTAGAGGGACTGGCGCGCCGGGTGGATTCTTTGGTCTCATTTTATAGCAAAAAAGCGCGGGTACAGAAGCGAGGAGCAAGATCA
TTGGTTAGTTTGAATAAAAACTCCGGTCACGAGAGGGTTCTAGAAACTGCTGAGGGGCACCCTAGAAACTGTTATGAGCTCAGCTATTTCTGGGA
ATTGTTTGCATAAGTCCGGGGTGGGGACTCAGGTGCATGACCACAGATATCCCGTTTGGCACTCCAAATCTTCCTTGTTTTTCTTGCTCTGACCT
TTTACTGTGCCTGGCTGCACCTGTCCCTAAACCTTGGGCCCCCCGGGTAGTTTCCATACTCCACGGAATGATTTCTGCCTCATGATTTCTGCCTC
CTTCACTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCGCTCCACACACACACTCTCTCTGTTCTTTATACAGAACACACACAAAAAAATAAAAAAA
AATGCCTTCGGTTCCACACGATCTCACTCTCTGAAGATCTGCAAATGAGGTTCTGATATGTTTTACAAGGAAGCGCTTCTAAGTCAACTGCGCTC
GAATGTCCTAGCCTCTCATGGGTAGTGGCCAGATAGCCATAAGAACATCTTGTGCTAGTCAGTTAGGCTGCATATGGGC
```

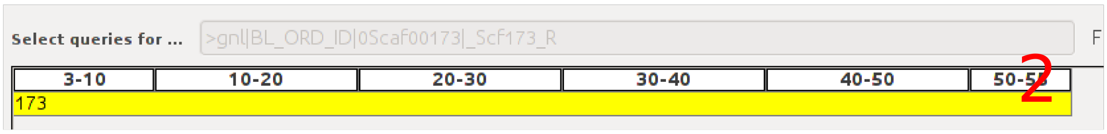
```
AG_Insertion_site_R
CGAGCTTACCATCCCGTCGGTGAGTTCAGGCGGTCCTTGGGTGGGAATCTCAGATCCCGGACGAGCCCCCAAATGAAAGACCCCAATGTTCGGGTAGTCCT
CCGACCTTGAGAAACCCTCCCAGGATCAACGCGAACACAACTCCGGATGCAAATAGCAAGAGGTTTATTGAACTTACGAGTACTCGGGCGACTCAGTCACC
TAGAGGGACTGGCGCGCCGGGTGGATTCTTTGGTCTCATTTTATAGCAAAAAAGCGCGGGTACAGAAGCGAGGAGCAAGATCATTGGTTAGTTTGAATAAA
AACTCCGGTCACGAGAGGGTTCTAGAAACTGCTGAGGGGCACCCTAGAAACTGTTATGAGCTCAGCTATTTCTGGGAATTGTTTGCATAAGTCCGGGGTGG
GGACTCAGGTGCATGACCACAGATATCCCGTTTGGCACTCCAAATCTTCCTTGTTTTTCTTGCTCTGACCTTTTACTGTGCCTGGCTGCACCTGTCCCTAA
ACCTTGGGCCCCCCGGGTAGTTTCCATACTCCACGGAATGATTTCTGCCTCATGATTTCTGCCTCCTTCACTCTCTCTCTCTCTCTCTCTCTCTCTCTCTC
GCTCCACACACACACTCTCTCTGTTCTTTATACAGAACACACACAAAAAAATAAAAAAAAATGCCTTCGGTTCCACACGATCTCACTCTCTGAAGATCTGC
AAATGAGGTTCTGATATGTTTTACAAGGAAGCGCTTCTAAGTCAACTGCGCTCGAATGTCCTAGCCTCTCATGGGTAGTGGCCAGATAGCCATAAGAACAT
CTTGTGCTAGTCAGTTAGGCTGCATATGGGC
```

Figure A.10: Scaf173 PCR Products Sanger Sequencing

Scaf00173

```
Scaf00173|oRV30ER
CAAGGGGACCTTATTTAATGCTAGGTCGTAGATTCAAACTAAAGCAAACAGACCCACATCCCCTGTTTCCTTATTTTTCTTTTTTAAGGCAAAAC
AAAAAAATAAAGTGGGGAAAGGAAAATACCAAAATAATTCATTAAAATTCAAATTCCCACCAGTCTTCTCGCTTTAGATGCTTCCAGTTCCCACA
CACGCCCAAT
```

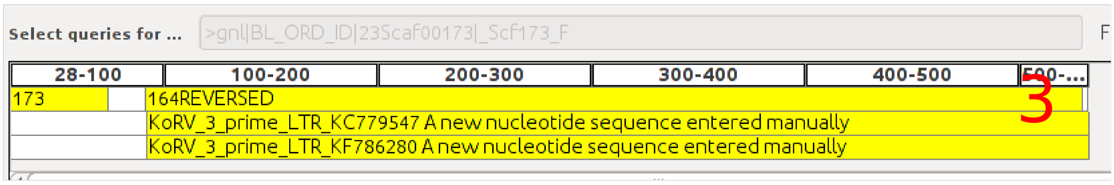
1



Select queries for ... >gnl|BL_ORD_ID|0Scaf00173|_Scf173_R

3-10	10-20	20-30	30-40	40-50	50-52
173					

2



Select queries for ... >gnl|BL_ORD_ID|23Scaf00173|_Scf173_F

28-100	100-200	200-300	300-400	400-500	500-...
173	164REVERSED				
	KoRV_3_prime_LTR_KC779547 A new nucleotide sequence entered manually				
	KoRV_3_prime_LTR_KF786280 A new nucleotide sequence entered manually				

3

```
Scaf00173|oRV30EF
ATATATATTCAACTGTAGCCCATACCCCTACTTGGAATGTTACCTTATTACAGGAATACTAGACGCTTAATTGCGAAAAACGGATTCTTAACTCA
CTGCAACGAGGGAGCCGTGGCGGAGTGTTGGGATTCTTCCGTCGACTCCGGCGAGGCGGCGCCTATGGTTTTTGGAACACGGTTGGG
```

4

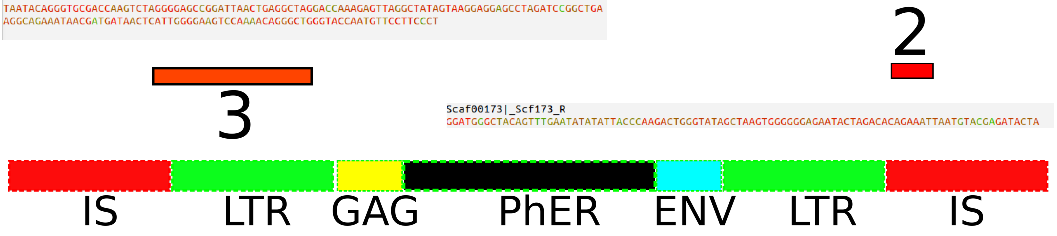
```
_Scaf173_R
ATCCTCGTTGTCTAGGGCTGGTACTTGTGCCCGAAACCCAGCCTAACAAAAAAAGGACGAGATCTGAAAATAATTAATGACGAAAATTCAGCCCGTAGAGG
GGAGGGTTCCCTATTGTTTTTGGGGGGGGAATCCCCCGATATTTTTTTTTCCGGGGGTCCCATCCCGTTACTTGATTGTTCCGGGGGCCTTTCCCAGAAT
GTTTTTGGGGGGGGGGGGTTGATCAAAATATACAAAGAAACTGGCCCCCCCCATTGGTTCCTCCCGCTTTTTTTGCAAAAAAGAGAGCCAGGAATCCACCC
GGACCCAATTCCCCCTGGTTGAATGAGTCGCCCGAGAACCCGAAAGCCAAAAAACCCCTTGTAATTTTATCCCGAATTGTGGTCCCCGTTCACCCGGGAAG
GTTTCCCAGGCCGGAGGACTACCCCAGCAATGGGGGCTTTTCTTTTGGGGCTCGGCGAGGGATTAAAATTCCCCCCCAAGACCGCCAAACCCCTGGGGAGG
TAGCTGCCCGGCACCCCCCCTTTTTCTTTTTTTTTCTTAATCCAAAAACT
```

```
oRV_30_E_F
GGGATCATTACTGAGTCACGGGATCACCTCTATTCTGCTCGTGATAGAGCACCCATAGAAAATGGGGAGGCAGAGGAGAAAGTGAGAAGAAACAACACAAA
AAGGGGAACCACTGTTTTAAAAAGGGCCGAGATTTCTTAAATTAGGGCGGGGGGGATCTCCGAAATGGGGGGTGAGGGGGAGCTGATAATGGGGCGATG
```

```
cf173_KoRVE30F
GAATAGCTCTGCGGTCTTGGGCAGAATATTCAACTGTAGCCCAAACCCCTACGTGGAGGTTACCTTATTACAGCAGCAAGCCAGAAAAATTGTGGAGGAGA
AGAATTTTATTGAAGAAATAGAGGTTGCTTATGGGCGTGGCGAGGCCAAGGTGAGGAATGTTGGCGAGCCAAAGGCTTCCAGGGCGTCGGC
```

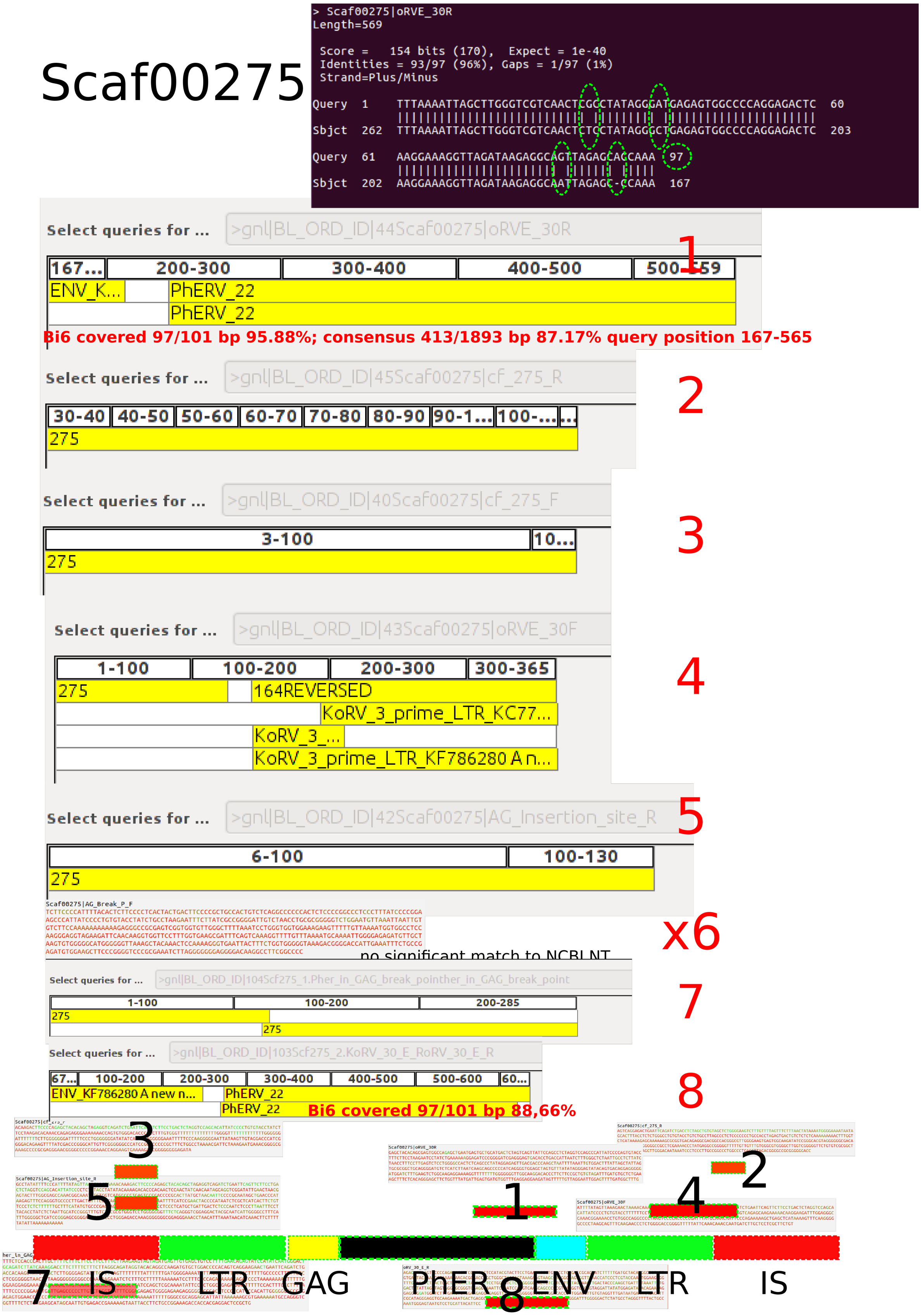
```
cf173_KoRVE30R
GAAAGGCCTCTTATCTCCTGTGGGTGATAGACTCAAACTATAGCCCACACCCCCCCATGCCCTTTTTCCTTATTTTCCTCTTCTAGGCCCAATTCTAATTG
TCGTGGGTGGAGGGGAAATGGGGGGAAAAAAAGGTTAATAACATTAAA
```

```
Scaf00173|_Scf173_F
ATTCAACTGTAGCCATACCCCTACTTGGTATGTTACCTTATTACAGGAATACTAGACACAGAAATTAATGTACGGGCCATCGTCACGCGGCCACC
TCCGGATGTGCATAGCAATGGGTTTATTGAACTTACGAGTACTCGGGCGACTCTACACCTAGAGGGACTGGCGCGCGCGTGGATTCTTTGGCCTC
TTGTTATATGCATCAAGCGTTGTGCAGAAGCGAGGAGTAAGATCATTGCTTAGTTTGAATAAAATATCCGGTCACTGGAGGCTTCTAGAAACTGC
TAAGGGGCTCCCGCAAAAACTGTTATGAGCTCAGCTATTTCTGGGAATTGTTTGCGTAAGTCCGGGGAGAGACTCAGGTGCATGACCACAAATAT
CCCGTTTGGCACTCCAAATCTTCCTTGTTTTTCTTGCTCTTGGCCTTTTACTGTGCCTGGTGCACCGGTGCCGCTAAAACCGTGTGCCCTCCGGG
TAGGCTTCCATACTCCACGGAATGGTTGCTGGCCTCAGTGATTTCAGCCTCCTTCATTCCCCCATTTTCTTTGGGGAACACGTCAAACTCTTGAC
GTGGTGGGTCTTGCTCATTATAAGTTGCAAAAGAAGGAAGACACCAGGATCCGATAACGTGGTACCTTTAACGCAACAGTTATCTTGAACTAATT
TAATACAGGGTGCGACCAAGTCTAGGGGAGCCGGATTAACTGAGGCTAGGACCAAAGAGTTAGGCTATAGTAAGGAGGAGCCTAGATCCGGCTGA
AGGCAGAAATAACGATGATAACTCATTGGGGAAGTCCAAAACAGGGCTGGGTACCAATGTTCCTTCCCT
```



```
Scaf00173|_Scf173_R
GGATGGGCTACAGTTTGAATATATATTACCCAAGACTGGGTATAGCTAAGTGGGGGAGAATACTAGACACAGAAATTAATGTACGAGATACTA
```

Figure A.11: Scaf275 PCR Products Sanger Sequencing



```
> Scaf00275|oRVE_30R
Length=569

 Score =   154 bits (170),  Expect = 1e-40
 Identities = 93/97 (96%), Gaps = 1/97 (1%)
 Strand=Plus/Minus

Query  1    TTTAAAATTAGCTTGGGTCGTCAACTCGCCTATAGGGATGAGAGTGGCCCCAGGAGACTC  60
            ||||||||||||||||||||||||||| ||||||||| ||||||||||||||||||||||
Sbjct  262  TTTAAAATTAGCTTGGGTCGTCAACTCTCCTATAGGGCTGAGAGTGGCCCCAGGAGACTC  203

Query  61   AAGGAAAGGTTAGATAAGAGGCAGTTAGAGCAGCAAA  97
            ||||||||||||||||||||||| ||||||| |||||
Sbjct  202  AAGGAAAGGTTAGATAAGAGGCAATTAGAGC-CCAAA  167
```

Bi6 covered 97/101 bp 95.88%; consensus 413/1893 bp 87.17% query position 167-565

Bi6 covered 97/101 bp 88,66%

```
Scaf00275|AG_Break_P_F
TCTTCCCCATTTTACACTCTTCCCCTCACTACTGACTTCCCCGCTGCCACTGTCTCAGGCCCCCCACTCTCCCCGGCCCTCCCTTTATCCCCGGA
AGCCCATTATCCCCTGTGTACCTATCTGCCTAAGAATTTCTTATCGCCGGGGATTGTCTAACCTGCGCGGGGGTCTGGAATGTTAAATTAATTGT
GTCTTCCAAAAAAAAAAAGAGGGCCGCGAGTCGGTGGTGTTGGGCTTTTAAATCCTGGGTGGTGGAAAGAAGTTTTTGTTAAAATGGTGGCCTCC
AAGGGAGGTAGAAGATTCAACAAGGTGGTTCCTTTGGTGAAGCGATTTCAGTCAAAGGTTTTGTTTAAAATGCAAAATTGGGGAGAGATGTTGCT
AAGTGTGGGGGCATGGGGGGTTAAAGCTACAAACTCCAAAAGGGTGAATTACTTTCTGGTGGGGGTAAAGACGGGGACCATTGAAATTTCTGCCG
AGATGTGGAAGCTTCCCGGGGTCCCGCGAAATCTTAGGGGGGGAGGGGACAAGGCCTTCGGCCCC
```

no significant match to NCBI NT

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig angefertigt, nicht anderweitig zu Prüfungszwecken vorgelegt und keine anderen als die angegebenen Hilfsmittel verwendet habe. Sämtliche wissentlich verwendete Textausschnitte, Zitate oder Inhalte anderer Verfasser wurden ausdrücklich als solche gekennzeichnet.

Berlin, den 03.03.2019

Ulrike Löber