


RESEARCH ARTICLE

DNA sonication inverse PCR for genome scale analysis of uncharacterized flanking sequences

David E. Alquezar-Planas^{1,2}  | Ulrike Löber^{1,3,4}  | Pin Cui¹ | Claudia Quedenau⁵ | Wei Chen⁶ | Alex D. Greenwood^{1,7} 

¹Department of Wildlife Diseases, Leibniz Institute for Zoo and Wildlife Research, Berlin, Germany; ²Australian Museum Research Institute, Australian Museum, Sydney, NSW, Australia; ³The Berlin Center for Genomics in Biodiversity Research, Berlin, Germany; ⁴Experimental and Clinical Research Center, A Cooperation of Charité – Universitätsmedizin Berlin and Max Delbrück Center for Molecular Medicine, Berlin, Germany; ⁵Genomics, Max Delbrück Center for Molecular Medicine, Berlin, Germany; ⁶Berlin Institute for Medical Systems Biology, Max-Delbrück Center for Molecular Medicine, Berlin, Germany and ⁷Department of Veterinary Medicine, Freie Universität Berlin, Berlin, Germany

Correspondence

David E. Alquezar-Planas
Email: david.alquezar@austmus.gov.au

Ulrike Löber
Email: ulrike.loeber@mdc-berlin.de

Alex D. Greenwood
Email: greenwood@izw-berlin.de

Present address

Ulrike Löber, Max Delbrück Center for Molecular Medicine, Host-microbiome factors in cardiovascular disease, Berlin, 13125, Germany

Pin Cui, ShenZhen HaploX Biotechnology Co, LTD, Nanshan District, Shenzhen, Guangdong 518057, China

Wei Chen, Department of Biology, Southern University of Science and Technology, Shenzhen, Guangdong 518055, China

Funding information

This work was supported by a postdoctoral scholarship from the Deutscher Akademischer Austauschdienst (DAAD; grant number 2014 57129705 to D.E.A.-P.) and the Morris Animal Foundation (grant number D14ZO-94 to A.D.G. & D.E.A.-P.).

Handling Editor: Susan Johnston

Abstract

1. There are few available tools to comprehensively and economically identify uncharacterized flanking regions that are not extremely labour intensive and which exploit the advantages of emerging long-read sequencing platforms.
2. We describe SIP; a sonication-based inverse PCR high-throughput sequencing strategy to investigate uncharacterized flanking region sequences, including those flanking mobile DNA. SIP combines unbiased fragmentation by sonication and target enrichment by coupling outward facing PCR priming with long-read sequencing technologies.
3. We demonstrate the effectiveness of SIP by determining retroviral integrations which are high copy and challenging to characterize. We further describe SIP's workflow, examine retroviral (proviral) enrichment and characterize viral structural variants identified. When SIP was coupled with long-read sequencing using the PacBio RS II platform, proviral integration was extensively characterized at high sequence depth per integration. By interrogating the sequence data, we were also able to test several intrinsic factors including SIP's propensity to form chimeric sequences and adapter ligation efficiencies.
4. SIP is an adaption of a traditional molecular biology technique that can be used to characterize any unknown genomic flanking sequence or to extend any sequence for which only minimal sequence information is available. SIP can be applied broadly to study complex biological systems such as mobile genetic elements with high throughput.

KEYWORDS

inverse PCR, KoRV, long-read sequencing, sonication, uncharacterized flanking regions

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint first authors.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2020 The Authors. *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of British Ecological Society

1 | INTRODUCTION

There are relatively few molecular methodologies for identifying sequences flanking a known sequence that have been adapted to current genomic sequencing approaches. Presently, various PCR-based strategies including linear amplification-mediated PCR (LAM-PCR; Schmidt et al., 2002, 2007), ligation-mediated PCR (LM-PCR) and splinkerette PCR (Devon et al., 1995) are the most exploited methods used. All three methods require the digestion of genomic DNA (gDNA) with a restriction endonuclease, the ligation of a linker cassette or adapter(s) and the amplification via primers that anneal to conserved regions of a known sequence and the ligated adapter(s). However, their reliance on restriction enzymes limits exhaustive characterization of flanking sites, particularly for repetitive or multi-copy sequences, 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; Gabriel et al., 2009) showed increased efficiency as compared to its predecessor and is currently one of the most robust approaches in use for flanking site retrieval, for example, for multi-copy integrated retroviruses (Giordano et al., 2015). However, given that it is based on standard PCR, it is limited to targeting one flanking site (5' or 3') at a time.

Mobile genetic elements, for example, retroelements, transposons and plasmids, represent classes of sequence, where defining flanking regions can be particularly difficult. While these elements share in their ability to change or increase the number of positions across a genome, they are classified by the chromosomal integration mechanisms employed to do so. The process of transposition is in all instances mutagenic, as the integration sites represent a permanent alteration of the host DNA within a cell. However, the extent and impact of these mobile elements on genome evolution, function and disease potential vary significantly across taxa and transposon class and remain subjects of ongoing investigation. The ability to characterize integration sites from mobile elements across a host genome is fundamental in their study, including determining their integration site preferences. Nevertheless, identifying integration sites comprehensively is generally challenging, as one must identify similar sequences integrated into hundreds to thousands of genomic locations. Shotgun sequencing approaches are feasible but are expensive, not amenable to large numbers of individuals and bioinformatically challenging when no prior enrichment for specific targets is applied.

Inverse PCR (Ochman et al., 1988) is a variant of PCR that has historically been used to obtain flanking sequences (Nowrouzi et al., 2006; Silver & Keerikatte, 1989). Its premise requires the fragmentation of genomic DNA (gDNA) followed by the intra-molecular circularization of DNA fragments. Inverted PCR primers designed end to end on conserved regions of a DNA sequence, such as retroviral long terminal repeats (LTR), are then used for targeted amplification of unknown flanking regions. Since its development, inverse PCR has fallen out of contemporary use. However, an adaptation of the method holds several benefits in characterizing unknown flanking regions, particularly when coupled with long-read high-throughput sequencing platforms

that offer improvements in characterizing genetic variation that is highly repetitive and complex in nature. We describe SIP—a sonication-based inverse PCR strategy, coupled with Pacific Biosciences PacBio RS II platform. We evaluated SIP as a tool for comprehensive flanking sequence retrieval in a high copy integration model using the endogenizing koala retrovirus (KoRV) as an example. In this context, we (a) employed sonication to randomly fragment DNA and avoid the use of biased-cutting restriction enzymes, (b) we tested and demonstrated adapter ligation deficiencies across DNA ligation experiments including those used in the generation of blunt end high-throughput sequencing DNA libraries and (c) 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 from mobile genetic elements. SIP will therefore have broad applications from clinical settings to molecular evolutionary analysis (Table 1).

2 | MATERIALS AND METHODS

The SIP protocol in this study employed five primary steps: (1) DNA was extracted, randomly fragmented and end repaired, (2) the end repaired DNA was then divided into two groups. One group had an adapter ligated before circularization (Adapter Ligation Group), while the other did not (Non-Adapter Ligation Group), (3) the two groups were circularized into closed circles, (4) long terminal repeat (LTR) and polymerase gene (*pol*) primed long fragment PCR was performed on the closed circles, (5) the resulting KoRV integration-enriched PCR products were built into PacBio DNA libraries, sequenced and analysed. 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 gDNA ScreenTapes.

1. DNA extraction, fragmentation and end repair

Genomic DNA was extracted using a standard silica-based tissue extraction kit, the QIAamp DNA Minikit (Qiagen). Products were fragmented using a Covaris ultrasonicator, which produced an average DNA fragment size of 2–7 kb in length. Sheared DNA was subsequently blunt-end repaired using the Fast DNA End Repair kit (Thermo Scientific) in triplicate.

2. Adapter ligation

Two complementary oligos were synthesized for adapter construction (Table S1). Each oligo contained a 5'-phosphate to facilitate subsequent blunt-end ligation. The oligos were annealed together by following the Illumina sequencing adapter preparation procedure (Meyer & Kircher, 2010). The Adapter Ligation Group was set up using a T4 DNA Ligase kit (5 U/ μ l; EL0014; Thermo Scientific) with 5 μ l of T4 DNA ligase buffer (10X), 5 μ l of 50% PEG 4000 solution,

TABLE 1 Clinical, taxonomic and alternative applications of SIP

Application	Description	Published studies
Health and disease		
Transposable elements	The simultaneous characterisation of transposable elements (e.g. integrating viruses) and their integration sites across multiple genomic locations	Löber et al. (2018)
Viral characterization	Characterisation of divergent viruses from conserved regions when limited reference data is available. This can be used to characterise both large DNA viruses (e.g. herpes) as well as RNA viruses through initial cDNA synthesis that are in low titre. This is particularly relevant from a viral metagenomic context when contaminating host DNA exceeds viral nucleic acids, making shotgun sequencing cost prohibitive	Geldenhuis et al. (2018)
Insertional mutagenesis	The identification of host gene disruption and oncogenesis	Ranzani et al. (2013)
Gene therapy	Viral vector integration used across gene therapy trials and gene editing technologies (e.g. CRISPR-Cas9) used to add, alter or remove genome sequences can be studied. Similarly aberrant integrations or edits across the genome and their precise locations can be determined	Hanlon et al. (2019)
Chromosomal rearrangements	Malignancies such as tumours and cancer can result in gene or chromosomal rearrangements	Merker et al. (2018)
Microorganism strain level detection and antibiotic resistance	Detection of specific microbes of interest from conserved genomic regions of a certain species or genus. Long reads might enable strain level detection. Targeting antibiotic resistance genes coupled with long-read sequencing is of relevance as well	Kim et al. (2016)
Non-invasive disease diagnostics	Identification of rare mutations from circulating free DNA in liquid biopsies (e.g. blood and pre-natal fluid) for genetic testing of disease. For example cancer, infection, hereditary disorders and transplant rejection	Bronkhorst et al. (2019)
Antibody discovery and engineering	Selection of phage/yeast display. Simultaneous resolution of multiple variable regions (VH/VL)	Ferrara et al. (2018)
Taxonomic classification		
Gene duplications or repetitive regions	Multiple copies of genes (e.g. gene duplications) or long repetitive regions such as tandem repeats could be studied. E.g. Many immune genes such as the Major Histocompatibility Complex (MHC), which are generally challenging to classify using standard PCR, Sanger sequencing and short read high-throughput sequencing platforms due to their inherent nature (e.g. inwards facing primers) and limited read length (~1,000 bp). Additionally, short read platforms are often unable to assemble repetitive regions	Trowsdale and Knight (2013)
Gene rearrangements	Gene re-arrangements across mitochondrial and nuclear genomes in non-model organisms could be investigated. This is particularly relevant when there is limited sequence information available across genera and/or families and an uncharacterised target species doesn't follow conventional annotation to available reference data. Examples of gene rearrangements have been identified across numerous wildlife lineages	Chen et al. (2018)
Resequencing	Resequencing of parts of the genome that have been assembled with short read sequencers. Or to verify gaps across assemblies due to low sequencing depth	Chaisson et al. (2015)
Environmental DNA	An alternative approach to sequence capture and metabarcoding for the detection of specific species across environmental samples	Seeber et al. (2019)
Population genetics	A comparable approach to a technique that targets transposable elements for genome wide SNP discovery and population genetics	Rey-Iglesia et al. (2019)
Other		
Molecular biology applications	Testing of molecular biology procedures. E.g. In this study, this method was used to determine ligation efficiency of blunt-end adapters	This study—Alquezar-Planas et al. (2020)
Functional studies — Promotor, enhancer detection, gene discovery and RNA expression	Upstream and downstream flanking regions of a target gene could be studied to identify or detect promotors and enhancers of that gene. Additionally, genes with unknown function within a flanking region could be investigated. Splice variants of relevant genes could also be examined through a targeted approach. This is relevant when the target is an expressed exon and additional exons are incorporated in the transcripts. This would be difficult to achieve using short-read sequencers as the various transcripts produced will all map to the same exons. SIP would be useful in studying expression at low titres	Symmons and Spitz (2013)

1 μ l of adapter (50 μ M), 2.5 μ l of T4 DNA ligase and 36.5 μ l of blunt-ended DNA in a 50- μ l 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). DNA size among the purified products from the Adapter Ligation Group showed a similar size distribution to the blunt-ended DNA from the Non-Adapter Ligation Group (Figure S1).

3. Circularization of fragmented DNA into closed circles

To find the optimal ligation conditions for subsequent inverse PCR, we performed a series of nine ligations using a gradient of (total) input blunt-ended DNA as previously described (Löber et al., 2018). Ligation reactions for both the Adapter Ligation Group and the Non-Adapter Ligation Group were set up using a commercially available T4 DNA Ligase kit (5 U/ μ l; Thermo Scientific). Ligation was performed in a thermal cycler at 16°C for 16 hr followed by enzyme inactivation at 70°C for 5 min. A non-template circularization control (control 1) was run simultaneously for each gradient. All ligations for both groups and control 1 were performed in triplicate and subsequently pooled.

4. Inverse PCR

Inverse PCR was performed as previously described (Löber et al., 2018). Briefly, KoRV proviral genomes were downloaded from GenBank (accessions: KF786280, gen, KF786282, KF786283, KF786284, KF786285, KF786286, AB721500, KC779547) and aligned using the MAFFT plug-in in Geneious (v7.1.7) using default settings. Inverse primers were designed to conserved regions on the KoRV LTR and to the middle of the polymerase gene (*pol*) using Primer3Plus software (Untergasser et al., 2012; Table S1). PCR was done using MyFi Mix (Bioline GmbH). A non-template PCR control (control 2) and a linear control of fragmented blunt-ended genomic koala DNA were included (control 3). A 40-ng input DNA (conc. 0.8 ng/ μ l in circularization) for both the Adapter Ligation Group and the Non-Adapter Ligation Group were chosen as the optimal circularization product-based TapeStation quantitation within the (a) 600–700 bp and (b) 2–7 kb range.

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 as previously described (Löber et al., 2018). The resulting four samples from these purifications were then prepared as sequencing libraries using the PacBio (Pacific Biosciences) 5-kb template prep protocol and the SMRTbell™ 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 2,100 Agilent Bioanalyzer and the 1,200 DNA chemistry (Agilent Technologies). Sequencing on the PacBio RS II platform was done as previously described (Löber et al., 2018), using the MagBead Standard protocol, C4 chemistry and P6 polymerase on a single v3 Single-Molecule Real-Time (SMRT) cell with 1 \times 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). The reads from the insert sequence were processed within the SMRT®Portal browser (minimum full pass = 1; and a minimum predicted accuracy of 90).

2.1 | Blunt-end SMRTbell DNA library construction and PacBio sequencing using Illumina adapters

The same fragmented koala extract above was used to build a blunt-end high-throughput sequencing DNA library. Library construction was based on the general principal developed for blunt-end library construction (Margulies et al., 2005), which was later modified using Illumina adapters and is one of the standard protocols for generating Illumina libraries (Meyer & Kircher, 2010). The following modifications were made to the previously described Illumina adaptation: (a) All SPRI bead purification steps were substituted with spin column modifications (QIAquick PCR purification kits; Qiagen), (b) a final adapter concentration of 1 μ M was used to build the libraries—the same concentration as the Adapter Ligation Group (c), the fill-in reaction procedure was performed at 65°C for 20 min, (d) all columns were incubated at 37°C for 5 min prior to elution and (e) the final purification following the fill-in reaction was omitted (Gansauge et al., 2017). 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/ μ l (40 ng total DNA input) was used as it was previously determined to be the optimal ligation concentration for the sample. The blunt-end DNA library with Illumina adapters was subsequently built into a PacBio sequencing library as described above.

2.2 | Bioinformatic analysis

2.2.1 | Calculating KoRV sequence enrichment

A KoRV reference database was created by downloading the genomes of KoRV-A (KF786280) which is known to endogenize in the koala genome and KoRV-B (KC779547) which is the exogenous KoRV variant with highest prevalence. All four datasets were searched for KoRV. All 'Reads of Insert' were aligned to the KoRV references using megablast (Altschul et al., 1990) with default settings.

KoRV positive Reads of Insert were aligned to the NCBI nt database (NCBI-GenBank Flat File Release 220.0) using megablast with

e-value restriction of 10^{-5} . Results were visualized using KronaTools (Ondov et al., 2011). The same alignment and visualization process was applied to sequences, determined as off-target reads, which could not be aligned to KoRV.

2.2.2 | Adapter search

All Reads of Insert were separately aligned to KoRV domains (LTR, *gag*, *pol*, *env*), primer sequences and adapter sequences (BLASTn). 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 4.

3 | RESULTS

3.1 | Development and testing of SIP

This study is a companion paper to a previously published study (Löber et al., 2018), in which SIP was applied to a zoo koala and compared to the koala reference genome to determine the number of shared versus unique integrations between the two koalas. The data generated have been deposited in the National Centre for Biotechnology Information (NCBI) Sequence Read Archive (SRA; <https://www.ncbi.nlm.nih.gov/sra/>; accession no. SRS2321692). This study focuses on SIP's effectiveness as a tool for unknown flanking sequence characterization. It also presents novel data on blunt-end adapter ligation efficiencies used in DNA library construction employed across various omics methods.

A visual summary of SIP is presented in Figure 1. The method requires the initial circularization of fragmented blunt-ended koala DNA, followed by targeted amplification of KoRV using primers to the long terminal repeat (LTR) and the polymerase gene (*pol*; Figure S1). SIP provides an alternative to previously described methods in a simplified workflow (Figure 2). This enables the dual characterization of a known sequence, for example, a mobile genetic element, and unknown flanking sequence on the studied genome through long-read high-throughput sequencing.

The development of an optimized workflow for SIP required the testing of several experimental conditions in order to (a) establish the optimal inverse PCR condition requirements and (b) implement controls. Pooled triplicates of circularized koala gDNA were used as template for SIP. TapeStation readings of inverse PCR 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 artefacts. We hypothesize this was due to over amplification of a reduced amount of starting template, which resulted in the formation of large DNA concatamers.

Three controls were established to monitor SIP's performance (Figure 2). 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 inverse PCR reaction for each gradient and group. TapeStation assessments of purified products from Controls 1 and 2 resulted in no visible amplification products, thereby confirming the absence of DNA contamination. Control 3 consisted of fragmented blunt-ended koala gDNA. TapeStation readings displayed some minor observable amplification peaks, suggesting that un-circularized (linear) DNA could be amplified with primers in the inverse orientation. Standard PCR amplification could occur if more than one LTR is located on the same fragmented DNA molecule, on either the same provirus or across different proviruses. Non-circularized DNA may also be primed by a single PCR primer to produce amplicon products through a linear (non-exponential) amplification (Figure S2).

3.2 | Library length distribution, KoRV sequence enrichment and off target enrichment

Central to SIP's application is the intra-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 1 and 2). The premise behind the Adapter Group was to mark the sheared boundaries of the blunt-ended DNA fragments, important for biological interpretation of inverse PCR products.

As an adapted inverse PCR technique, it was initially unclear whether intra-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 size selected (described as long and short insert libraries—refer to Table 2 and Figure S3) using two different length cut-offs (refer to methods for details), and to compare 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 compared to off-target non-KoRV reads. Especially notable were the non-adapter long and non-adapter short datasets, which yielded total KoRV enrichment rates of 94% and 95% of all sequenced reads respectively (Table 3). In contrast, the adapter (long and short) datasets had a lower total enrichment rate of 82% and 63% respectively. The highest KoRV enrichment for sequences longer than 1,000 bp derived from the two long insert libraries at 96% (non-adapter long) and 97% (adapter long). While the shorter datasets displayed a reduced enrichment of 58% (non-adapter short) and 77% (adapter

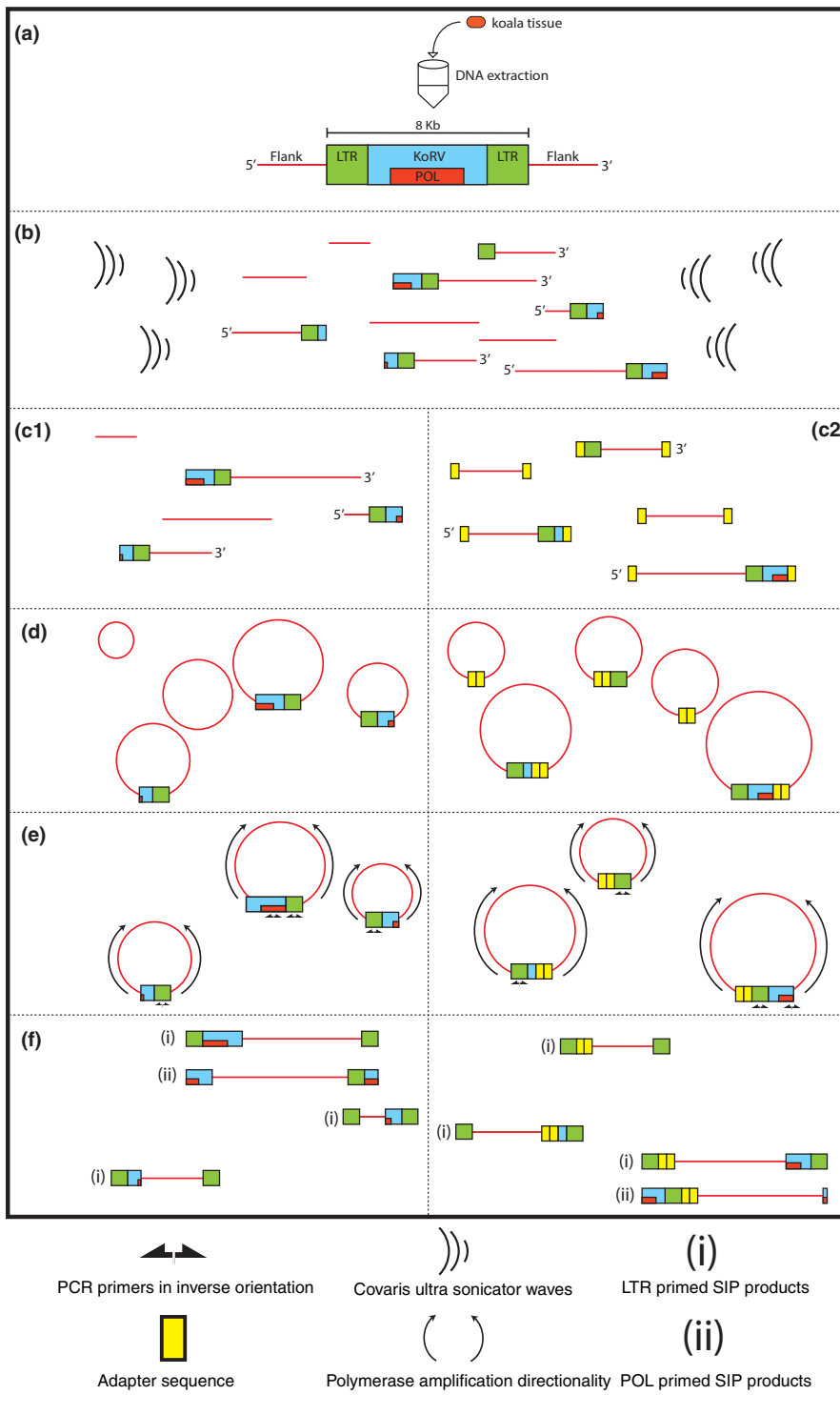


FIGURE 1 Visual representation of Sonication Inverse PCR (SIP). Abbreviations used in the figure include KoRV—koala retrovirus, LTR—long terminal repeat, *pol*—polymerase gene. (a) The KoRV provirus, which is integrated into the koala genomic DNA, is illustrated with typical LTR regions (green box) flanking the retroviral genes (blue box). Note: Only the approximate location of the *pol* gene (red box) is represented diagrammatically for simplicity. (b) Koala genomic DNA was fragmented to an average length of 2–7 kb using ultrasonication. The fragmented DNA was then blunt-end repaired and phosphorylated (not depicted). (c) The sample was subsequently divided in two; a Non-Adaptor Group (c1) and an Adaptor Group (c2). The Non-Adaptor Group was not modified in any way prior to circularization, while the Adaptor Group had an identical adaptor 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 Adaptor and Non-Adaptor 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 and sequenced products were inverted with the primer-binding site located on the flanks of the amplicon. Two primary types of PCR product were generated: (i) PCR products amplified by the LTR primers and (ii) PCR products amplified by the *pol* primer

short) respectively. The enrichment of KoRV sequences across the four datasets exhibited a mean alignment length between 1,111 and 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 intracircularization process can readily produce sequenceable amplicons of interest nearing 10 kb in length.

A breakdown of the off-target (non-KoRV) sequences (Supplemental Files 1–4) visualized using Krona (Ondov et al., 2011) showed that from

51% to 68% of the non-KoRV sequences from the four datasets showed high similarity to the Tammar wallaby *Notamacropus eugenii* (Renfree et al., 2011); the closest related species to the koala with an assembled reference genome. The second largest fractions (16%–28%) matched the koala genome *Phascolarctos cinereus* (Johnson et al., 2018). Overall, between 79% and 84% of off-target reads were similar to wallaby or koala sequences. In addition, approximately 6%–10% of the sequences could be assigned to other eukaryotes, notably extant marsupials such as the Tasmanian devil, platypus and opossum; while only a fraction of

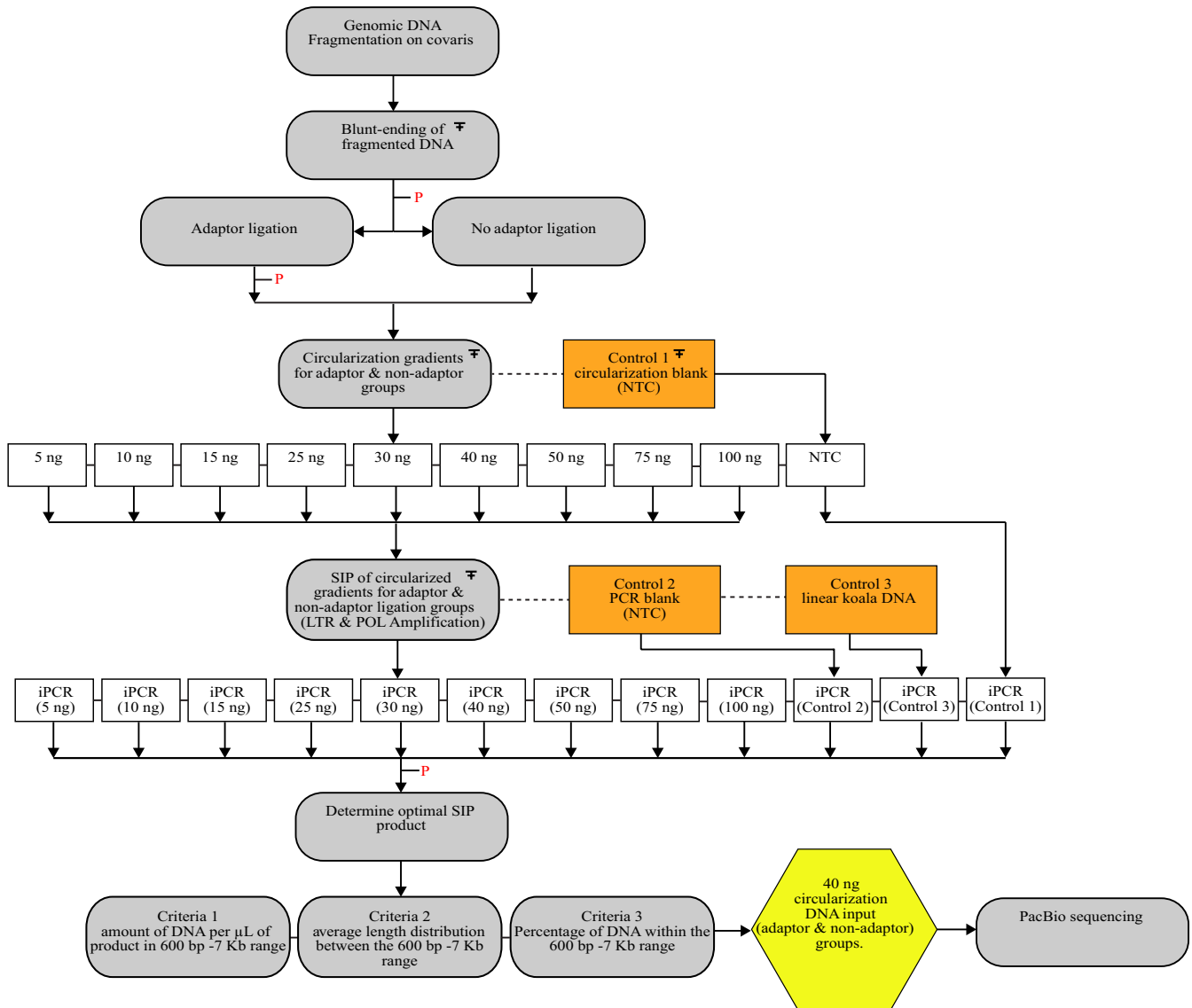


FIGURE 2 Experimental workflow of Sonication Inverse PCR (SIP) methodology. Abbreviations used in the figure include SIP—Sonication Inverse PCR, P—purification, T—Triplicate reactions, NTC—non-template control, iPCR—inverse Polymerase Chain Reaction. Grey-rounded rectangular boxes denote important steps in the workflow, white rectangles represent 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 Adaptor Group, where an adaptor was ligated on either end of the DNA fragment pool, or a Non-Adaptor 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 the 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 inverse PCR 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

reads (0%–0.9%) across the four datasets could not be matched to any public nucleotide sequence from NCBI. The high amount of off-target reads matching genomic DNA from other marsupials than the koala might result from the number of sequences represented in the NCBI nt database. There are approximately 10-fold more wallaby sequences deposited in GenBank than koala sequences despite draft genomes

for both being deposited and described (Johnson et al., 2018; Renfree et al., 2011). Despite a search against the entire NCBI nucleotide database, the analysis of the reads yielded no identifiable bacterial sequences. A re-analysis of the off-target sequences at the protein level displayed a comparable result to that of the nucleotide analysis (data not shown).

Dataset	Library information as per bioanalyzer	Sequence read information from each library		
	Average length	Minimum length	Maximum length	Average length
Non-adapter (long insert library)	3,500	16	9,591	2,348
Non-adapter (short insert library)	1,600	15	6,632	1,083
Adapter (long insert library)	3,500	16	9,865	2,176
Adapter (short insert library)	1,600	16	6,293	1,256

TABLE 2 General DNA library information**TABLE 3** KoRV sequence enrichment using SIP and PacBio RS II sequencing

Dataset	Total no. reads	Number of KoRV-like reads identified			KoRV enrichment		KoRV alignment length	
		Total	<100 bp	>1,000 bp	% KoRV-like enrichment (total)	% KoRV-like enrichment (>1,000 bp)	Mean	Maximum
Non-adapter (long insert library)	28,983	27,367	136	26,368	94	96	2,396	9,590
Non-adapter (short insert library)	31,794	30,054	365	17,540	95	58	1,111	4,972
Adapter (long insert library)	24,076	19,663	91	19,118	82	97	2,302	9,864
Adapter (short insert library)	26,910	16,841	64	13,021	63	77	1,321	6,292

TABLE 4 Adapter counts across four standard SIP datasets and one blunt-end high-throughput DNA library adapter dataset

Dataset	Total filtered reads ^a	Filtered reads with two adapters		Filtered reads with >2 adapters		Filtered reads with no adapters		Filtered reads with other configurations ^d	
		Total ^b	Percentage ^c	Total ^b	Percentage ^c	Total ^b	Percentage ^c	Total ^b	Percentage ^c
Non-adapter (long insert library)	27,060	0	0	0	0	27,060	100	0	0
Non-adapter (short insert library)	28,329	0	0	0	0	28,329	100	0	0
Adapter (long insert library)	22,730	455	2	270	1.19	21,818	95.99	187	0.82
Adapter (short insert library)	25,600	182	0.71	244	0.95	24,878	97.18	296	1.16
Blunt-end DNA library	18,327	4,426	24.2	562	3.07	10,745	58.63	2,584	14.1

^aTotal number of reads that passed filtering criteria.

^bTotal number of filtered reads with adapters or no adapter sequences.

^cPercentage of reads normalized to 10,000 with adapters or no adapter sequences.

^dTotal filtered reads with other adapter configurations. For example reads with two identical adapters

3.3 | Blunt-end adapter ligation efficiency and blunt-end 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

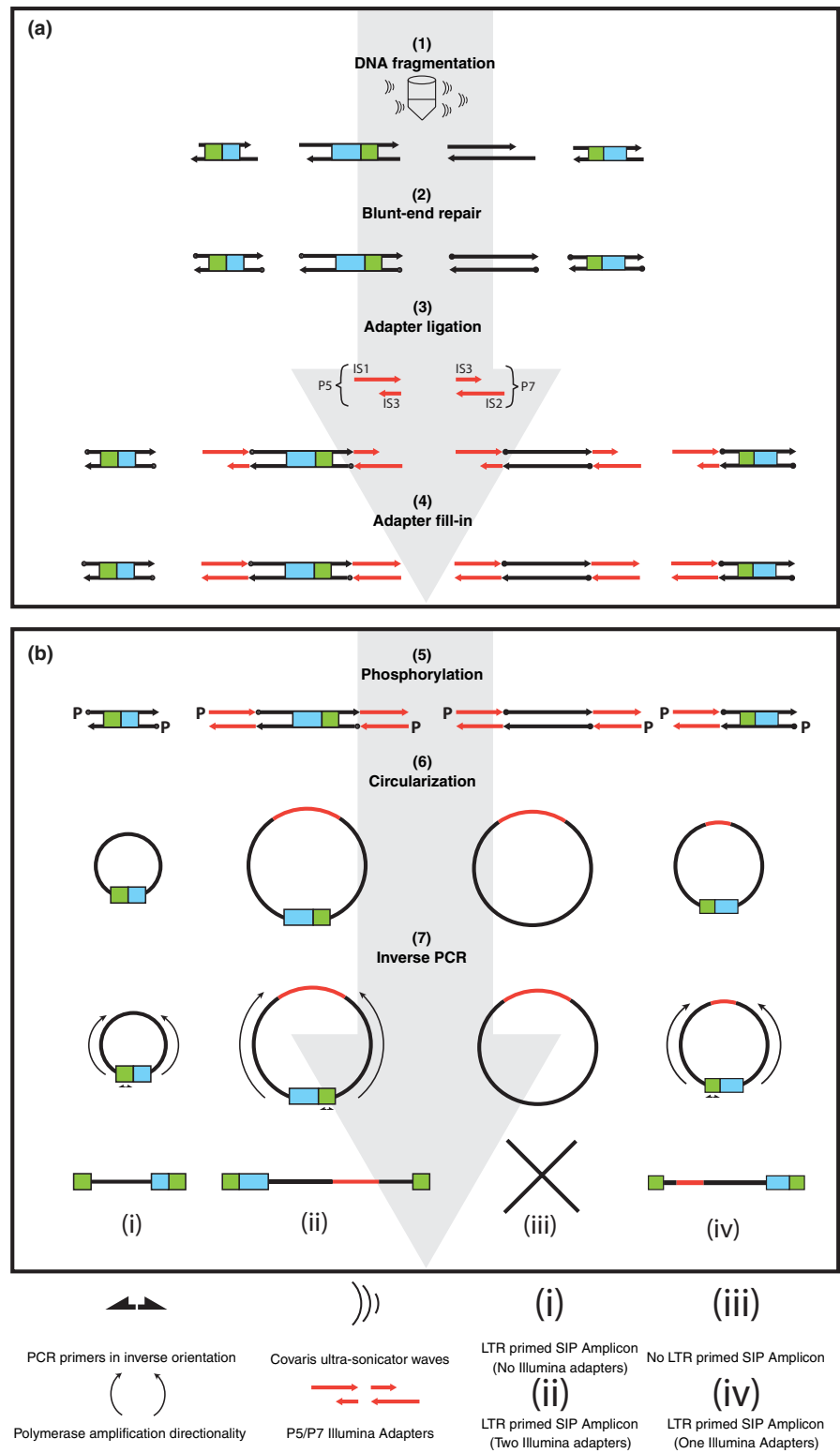
datasets without adapters. However, adapter ligation enrichment for the two datasets with adapters was low, with the highest percentage of filtered reads with two adapters occurring in the adapter long dataset (4%), while the adapter short dataset had an adapter ligation efficiency of 2%. Therefore, the majority of reads (approximately 86%–90%) in both standard adapter SIP datasets did not contain any adapter sequences (Table 4).

The data from the two standard SIP datasets with the incorporation of an adapter (long and short insert libraries) suggest 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) through creation of a Single Molecule Real-Time ('SMRTbell') DNA library. A blunt-end DNA library was generated from sheared koala gDNA using a variation of the Meyer & Kircher

protocol (Meyer & Kircher, 2010). The blunt-end DNA library was then subjected to the same circularization and KoRV LTR inverse PCR priming procedures as previously described. Importantly, the experimental approach amplifies circularized KoRV-blunt-end-DNA-library template regardless of whether Illumina blunt-end adapters are ligated to the ends of the DNA molecule (Figure 3). By priming the PCR in the KoRV LTRs, a comparative count of the

FIGURE 3 Blunt-end DNA library adaptor ligation experimental workflow.

(a) Blunt-end DNA library build: (1) Genomic koala DNA, represented by the black lines with arrows, was sonicated to an average length of 3–4 kb. The KoRV provirus (green box represents the LTR region, while the blue box represents 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 adaptors, Illumina P5 and P7, constructed from two different oligonucleotides (IS1 and IS3 or IS2 and IS3) were ligated to the ends of the blunted molecules. Adaptor ligation was not completely efficient and resulted in DNA templates with either none, one or two adaptors ligated to each DNA molecule. (4) A fill-in reaction repaired nicks and filled in the lagging adaptor strands. (b) SIP Procedure: (5) The DNA library was phosphorylated (indicated by a P), which added 5' phosphate and 3' hydroxyl groups. (6) Inter-molecular circularization of the DNA library ensues. Note: The circles represent double-stranded DNA. The reaction will circularize all DNA library molecules, regardless of the number of adaptors ligated to the distal ends (adaptors are denoted by a red line within the circle). (7) The circularized library was amplified using inverted KoRV LTR primers. Only circularized template with a KoRV LTR could amplify, irrespective of whether P5/P7 adaptors were ligated at the ends of the DNA libraries. LTR primed amplicons were then built into SMRTbell DNA libraries and sequenced on the PacBio RS II platform (not depicted). The numbers of reads with and without attached P5/P7 adaptors that were LTR primed were then informatically counted



enriched KoRV DNA molecules with and without Illumina P5 and P7 adapters attached could be calculated. Following, the blunt-end DNA library was subsequently converted into a SMRTbell DNA library, thereby incorporating both blunt-end Illumina adaptors and SMRTbell adapters, for sequencing on the PacBio RS II platform. An analysis of the sequence data from the blunt-end-SMRTbell DNA

library dataset indicated that only 24% of the reads had two adapter sequences in the correct orientation (Table 4). In contrast, 59% of reads had no adapters, 3% of reads had more than two adapters and 14% had the same adapter attached. Overall, approximately 76% of the DNA within a blunt-end DNA library is therefore not sequenceable; lacking the primer binding site for sequencing due to

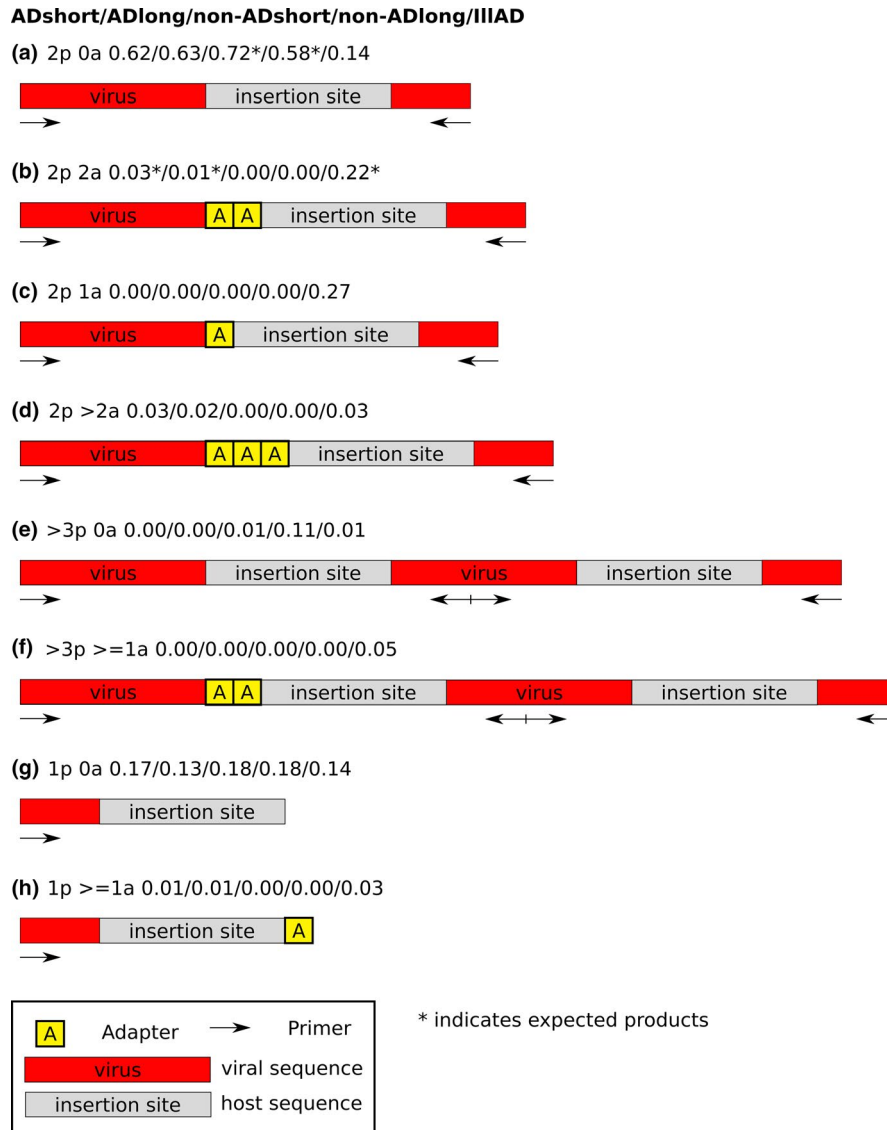


FIGURE 4 Identified structural variants of SIP sequences. Structural sequence variants identified in five SIP datasets are shown (structures a–h). The percentage of reads, where 1.00 = 100%, from each structure and for each dataset is indicated from left to right (adapter long, adapter short, non-adapter long, non-adapter short and the blunt-end SMRTbell 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-SMRTbell DNA library (Illumina adapter) ligation experiments. The displayed structures represent 87% of the structures observed for the adapter long dataset and, respectively, 81% for adapter short, 91% for non-adapter long, 88% for non-adapter short and 89% for blunt-end-SMRTbell DNA library (Illumina adapter) dataset. (a) Two primers (either *pol* or LTR paired) were present at the end of the read; no adapter was incorporated, (b) two primers (either *pol* or LTR paired) were present at the end of the read; two adapters were incorporated, (c) two primers (either *pol* or LTR paired) were present at the end of the read; one adapter was incorporated, (d) two primers (either *pol* or LTR paired) were present at the end of the read; more than two adapters were incorporated, (e) two primers (either *pol* or LTR paired) were present at the end of the read; in addition, more than one internal primer was detected. Such structures were concatemers; with no adapter incorporated, (f) two primers (either *pol* or LTR paired) were present at the end of the read; in addition, more than one internal primer was detected. Such structures were concatemers; with at least one adapter incorporated, (g) only one primer was detected at either end of the read representing linear products; no adapter incorporated, (h) only one primer was detected at either end of the read, representing linear products, with at least one adapter incorporated

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.

3.4 | SIP structure variations

Bioinformatic analysis of the sequence reads from the five datasets (adaptor and non-adaptor—long and short datasets and the blunt-end-SMRTbell DNA library dataset) revealed eight different DNA sequence structures (Figure 4). Structure A, containing two primer sequences (either LTR or *pol*) and no adapters, was the most frequent across all but the blunt-end-SMRTbell DNA library dataset. This result is not unexpected given the reduced ligation efficiency described. We also 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 4 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 of the three datasets (non-adaptor long and the blunt-end-SMRTbell DNA library dataset) without an incorporated adapter (Structure E—Figure 4) contained 1% chimeric sequences, while the non-adaptor short dataset had a maximum of 11% chimeric sequences. In contrast, chimeric sequences with an adapter sequence ligated (Structure F—Figure 4) were only identified in 5% of sequences within the blunt-end-SMRTbell 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 4 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-adaptor long and non-adaptor short datasets (18%). Overall, the sum of linear sequences without and with an adapter sequence (Figure 4 Structures G and H) across the four datasets was between 14% and 18%. The results indicate that circularization of the fragmented DNA and subsequent inverse PCR was efficient enabling the preferential (exponential) amplification of circularized DNA versus (non-exponential) linear DNA.

4 | DISCUSSION

Starting from a limited amount of known sequence to identifying the sequences flanking it, is a challenge relevant to many analyses (Table 1). One common application is the identification of viral and mobile element (transposons, retrotransposons) integration sites across a host genome, which is central to understanding integration preferences and the biological effects of such integrations. This is particularly important when processing multiple samples in parallel where lack of prior enrichment would drastically increase the subsequent bioinformatic analysis. While several relevant

short-read high-throughput molecular techniques exist to study these processes (Brotherton et al., 2008; Gabriel et al., 2009; Huang et al., 2009; Maricic et al., 2010; Schmidt et al., 2002, 2007; Uren et al., 2009), an adaptation of long-read inverse PCR holds several benefits. 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 considerably across different library preparation methods. A recent study reported eight of nine commercially available library kits had a maximum adapter ligation efficiency no greater than 30% (Aigrain et al., 2016). Our result on blunt-end ligation efficiency, including those used for generating a blunt-end DNA library (adaptable across different sequencing platforms), further exemplifies the limitations of these processes. The blunt-end DNA library method (with the Meyer & Kircher library modifications; Margulies et al., 2005; Meyer & Kircher, 2010) is effective, inexpensive and commonly used across genomics studies. However, the low adapter ligation efficiency observed and the halving of useable molecules due to non-directional ligation of identical adapters will likely impair the final complexity of DNA libraries. This is particularly important in studies that use degraded or low amounts of template material (e.g. DNA from historical museum specimens or from environmental samples), as the sequence data recovered will not reflect the full diversity within a biological sample. Future experiments should examine and compare other adapter ligation techniques such as those used in ssDNA library construction (Gansauge et al., 2017; Gansauge & Meyer, 2013) before performing the intracircularization 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 diversity reported is not a true reflection of the full diversity within the sample. The analysis of SIP structure variants (Figure 4) indicates that the vast majority of the reads across the four datasets (82%–86%) were inverted. While this suggests that the majority of the DNA has been circularized, the exponential nature of PCR may have masked 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 unknown flanking integration sites. Notwithstanding, SIP yields data of biological relevance as we

recently demonstrated by extensively characterizing KoRV integration sites and comparing the results to the koala reference genome, identifying novel recombinant KoRVs (recKoRVs; Hobbs et al., 2017; Johnson et al., 2018; Löber et al., 2018). 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 of the primer binding site and the frequency of the target sequence being amplified, particularly for long fragment amplifications. The inverse orientation of the PCR primers designed on the LTR of a retrotransposon or provirus enables concurrent retrieval of 5' and 3' integration sites, which could be adapted to any sequence. Unlike standard PCR methods employed to study mobile genetic elements and their flanking integration sites, in SIP, both forward and reverse primers can be anchored in proximity to each other thereby eliminating the need to prime the reaction inwards from a ligated adapter. Our experiments displayed a total KoRV enrichment rate between 63% and 95% for our four standard SIP datasets. Given that these libraries were built from the same sample and that PCRs were performed in triplicate, it is unlikely that the differences exhibited are a reflection of the variability across the PCR assays, but rather the varied workflow as it relates to the incorporation or exclusion of an adapter 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 (Table 2). While there is likely an upper limit to the size of circularized products due to steric influences, the analysis of the SIP sequences revealed that fragments nearing 10 kb in length (9,865 bp) were successfully circularized, amplified and sequenced. This suggests that with a varied library preparation procedure and the progression of long-read platforms (e.g. PacBio Sequel System and Oxford Nanopore MinION), even larger fragments could be enriched and sequenced.

While the repeated LTR region further complicates assembling proviral structure from short reads, SIP introduces its own unique challenges. Unfortunately, given the adapter ligation process was inefficient, restructuring the rearranged inverse PCR sequences proved challenging. No approach 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. 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. This simplified re-orientation of the reads compared to data from short-read sequencers.

Despite the challenges listed above, our experiments demonstrate that SIP is a simple, robust and efficient methodology for the analysis of proviral integration sites. While this represents a common application, the methodology can be used broadly to characterize any unknown sequence flanking a known sequence (Table 1). Clinical applications requiring the identification of insertions/deletions or mutations across multiple genomic locations

are likely to benefit from SIP. In particular, the insertional mutagenic properties of transposable elements, gene editing technologies and chromosomal rearrangements caused by malignancies could be investigated (Merker et al., 2018). Given the long-read length, genes, promoters and enhancers located hundreds to thousands of bases from an integration can be studied in detail (Bradner et al., 2017). Multiple individuals can be enriched and sequenced at high depth efficiently and economically. The analysis of anti-microbial resistance genes in bacterial genomes, particularly those representing a small fraction of the microbiome and hence, potentially difficult to detect by shotgun sequencing, represents another clinical application. Similarly, SIP could also be applied to assist with the assembly and annotation of divergent viruses, many of which have genomes within the range of PacBio sequencing read lengths (Geldenhuis et al., 2018), as well as strain level detection across bacteria and parasites (Kim et al., 2016).

Bioinformatic approaches to detect mobile genetic elements using short-read sequencing data are based on resequencing, reference genomes and/or target libraries (Ewing, 2015). However, despite the use of sophisticated bioinformatics software, it is often impossible to map and even assemble short reads originating from genomic regions containing structural variation, repetitive sequences and high homology (Mantere et al., 2019). In these contexts, the classification of diverse sequences, poorly characterized genomic loci and non-model organisms represent additional utilities for SIP. While the human genome is arguably the most complete mammalian reference genome, these limitations are exemplified here, as previous assemblies have been found to contain numerous large gaps (Chaisson et al., 2015). Furthermore, structural variations such as indels, duplications, inversions and tandem repeats remain poorly understood due to the technical limitations of the tools used to target and study them. Several of these difficult to target regions can cause a range of Mendelian diseases and can be resolved using SIP or other targeted long-read sequencing applications (Wang et al., 2015). Taxonomic and phylogenetic classification of non-model species is hindered by limited representation in public sequence databases, many of which are only represented by single gene markers. While this is slowly changing, genomic re-arrangements (e.g. in mtDNA genomes) that do not follow conventional annotation are not uncommon across various wildlife lineages. SIP could be applied in the resolution of these structural anomalies, thereby assisting in correctly assigning phylogenetic relationships among related species (Chen et al., 2018). In like manner, population genetic studies as described by a similar study targeting transposable elements is also likely to be relevant (Rey-Iglesia et al., 2019). SIP is therefore expected to assist broadly across a range of genomic studies and biological disciplines.

ACKNOWLEDGEMENTS

The authors thank Wei Sun for fruitful discussions. We thank Hanna Vielgrader of the Tiergarten Schönbrunn, in Vienna for a very productive collaboration and access to samples from Bilyarra.

CONFLICT OF INTEREST

None declared.

AUTHORS' CONTRIBUTIONS

D.E.A.-P., P.C., W.C. and A.D.G. conceived the ideas and designed the methodology; D.E.A.-P., P.C. and C.Q. performed research; D.E.A.-P., U.L. and A.D.G. analysed data; D.E.A.-P., U.L. and A.D.G. wrote the paper.

ACCESSION NUMBERS

CCS has been deposited with the study accession SRP110681 and sample number SRS2321692: Koala_Bilyarra_ERV.

PEER REVIEW

The peer review history for this article is available at <https://publons.com/publon/10.1111/2041-210X.13497>.

DATA AVAILABILITY STATEMENT

The data supporting this article have been deposited in the National Centre for Biotechnology Information (NCBI) Sequence Read Archive (SRA; <https://www.ncbi.nlm.nih.gov/sra/SRS2321692>).

ORCID

David E. Alquezar-Planas  <https://orcid.org/0000-0001-5360-5263>

Ulrike Löber  <https://orcid.org/0000-0001-7468-9531>

Alex D. Greenwood  <https://orcid.org/0000-0002-8249-1565>

REFERENCES

- Aigrain, L., Gu, Y., & Quail, M. A. (2016). 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*, *17*, 458. <https://doi.org/10.1186/s12864-016-2757-4>
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
- Bradner, J. E., Hnisz, D., & Young, R. A. (2017). Transcriptional addiction in cancer. *Cell*, *168*(4), 629–643. <https://doi.org/10.1016/j.cell.2016.12.013>
- Bronkhorst, A. J., Ungerer, V., & Holdenrieder, S. (2019). The emerging role of cell-free DNA as a molecular marker for cancer management. *Biomolecular Detection and Quantification*, *17*, 100087. <https://doi.org/10.1016/j.bdq.2019.100087>
- Brotherton, P., Endicott, P., Beaumont, M., Barnett, R., Austin, J., Cooper, A., & Sanchez, J. J. (2008). Single primer extension (SPEX) amplification to accurately genotype highly damaged DNA templates. *Forensic Science International: Genetics Supplement Series*, *1*(1), 19–21. <https://doi.org/10.1016/j.fsigss.2007.10.111>
- Chaisson, M. J. P., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., & Eichler, E. E. (2015). Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, *517*(7536), 608–611. <https://doi.org/10.1038/nature13907>
- Chen, L., Chen, P.-Y., Xue, X.-F., Hua, H.-Q., Li, Y.-X., Zhang, F., & Wei, S.-J. (2018). Extensive gene rearrangements in the mitochondrial genomes of two egg parasitoids, *Trichogramma japonicum* and *Trichogramma ostrinae* (Hymenoptera: Chalcidoidea: Trichogrammatidae). *Scientific Reports*, *8*(1), 1–11. <https://doi.org/10.1038/s41598-018-25338-3>
- Devon, R. S., Porteous, D. J., & Brookes, A. J. (1995). Splinkerettes—Improved vectorettes for greater efficiency in PCR walking. *Nucleic Acids Research*, *23*(9), 1644–1645. <https://doi.org/10.1093/nar/23.9.1644>
- Ewing, A. D. (2015). Transposable element detection from whole genome sequence data. *Mobile DNA*, *6*(1), 24. <https://doi.org/10.1186/s13100-015-0055-3>
- Ferrara, F., Bradbury, A. R. M., & D'Angelo, S. (2018). Primer Design and Inverse PCR on Yeast Display Antibody Selection Outputs. In *Schizosaccharomyces pombe Methods and Protocols*, Methods in Molecular Biology (Vol. 1721, pp. 35–45). Humana Press. https://link.springer.com/protocol/10.1007%2F978-1-4939-7546-4_4
- Gabriel, R., Eckenberg, R., Paruzynski, A., Bartholomae, C. C., Nowrouzi, A., Arens, A., Howe, S. J., Recchia, A., Cattoglio, C., Wang, W., Faber, K., Schwarzwaelder, K., Kirsten, R., Deichmann, A., Ball, C. R., Balaggan, K. S., Yáñez-Muñoz, R. J., Ali, R. R., Gaspar, H. B., ... Schmidt, M. (2009). Comprehensive genomic access to vector integration in clinical gene therapy. *Nature Medicine*, *15*(12), 1431–1436. <https://doi.org/10.1038/nm.2057>
- Gansauge, M.-T., Gerber, T., Glocke, I., Korlevic, P., Lippik, L., Nagel, S., & Meyer, M. (2017). Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. *Nucleic Acids Research*, *45*(10), e79. <https://doi.org/10.1093/nar/gkx033>
- Gansauge, M.-T., & Meyer, M. (2013). Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nature Protocols*, *8*(4), 737–748. <https://doi.org/10.1038/nprot.2013.038>
- Geldenhuys, M., Mortlock, M., Weyer, J., Bezuidt, O., Seemark, E. C. J., Kearney, T., Gleasner, C., Erkkila, T. H., Cui, H., & Markotter, W. (2018). A metagenomic viral discovery approach identifies potential zoonotic and novel mammalian viruses in Neoromicia bats within South Africa. *PLoS ONE*, *13*(3), e0194527. <https://doi.org/10.1371/journal.pone.0194527>
- Giordano, F. A., Appelt, J.-U., Link, B., Gerdes, S., Lehrer, C., Scholz, S., Paruzynski, A., Roeder, I., Wenz, F., Glimm, H., von Kalle, C., Grez, M., Schmidt, M., & Laufs, S. (2015). High-throughput monitoring of integration site clonality in preclinical and clinical gene therapy studies. *Molecular Therapy – Methods & Clinical Development*, *2*, 14061. <https://doi.org/10.1038/mtm.2014.61>
- Hanlon, K. S., Kleinstiver, B. P., Garcia, S. P., Zaborowski, M. P., Volak, A., Spirig, S. E., Muller, A., Sousa, A. A., Tsai, S. Q., Bengtsson, N. E., Lööf, C., Ingelsson, M., Chamberlain, J. S., Corey, D. P., Aryee, M. J., Joung, J. K., Breakefield, X. O., Maguire, C. A., & György, B. (2019). High levels of AAV vector integration into CRISPR-induced DNA breaks. *Nature Communications*, *10*(1). <https://doi.org/10.1038/s41467-019-12449-2>
- Hobbs, M., King, A., Salinas, R., Chen, Z., Tsangaras, K., Greenwood, A. D., Johnson, R. N., Belov, K., Wilkins, M. R., & Timms, P. (2017). Long-read genome sequence assembly provides insight into ongoing retroviral invasion of the koala germline. *Scientific Reports*, *7*, 1–9. <https://doi.org/10.1038/s41598-017-16171-1>
- Huang, A. M., Rehm, E. J., & Rubin, G. M. (2009). Recovery of DNA sequences flanking P-element insertions in *Drosophila*: Inverse PCR and plasmid rescue. *Cold Spring Harbor Protocols*, *2009*(4), pdb.prot5199. <https://doi.org/10.1101/pdb.prot5199>
- Johnson, R. N., O'Meally, D., Chen, Z., Etherington, G. J., Ho, S. Y. W., Nash, W. J., Grueber, C. E., Cheng, Y., Whittington, C. M., Dennison, S., Peel, E., Haerty, W., O'Neill, R. J., Colgan, D., Russell, T. L., Alquezar-Planas, D. E., Attenbrow, V., Bragg, J. G., Brandies, P. A., ... Belov, K. (2018). Adaptation and conservation insights from the koala genome. *Nature Genetics*, *50*(8), 1102–1111. <https://doi.org/10.1038/s41588-018-0153-5>
- Kim, S., Park, Y.-J., & Kim, J. (2016). Inverse PCR for subtyping of *Acinetobacter baumannii* carrying ISAba1. *Journal of Microbiology*, *54*(5), 376–380. <https://doi.org/10.1007/s12275-016-6038-3>
- Löber, U., Hobbs, M., Dayaram, A., Tsangaras, K., Jones, K., Alquezar-Planas, D. E., Ishida, Y., Meers, J., Mayer, J., Quedenau, C., Chen, W., Johnson, R. N., Timms, P., Young, P. R., Roca, A. L., & Greenwood, A. D. (2018).

- Degradation and remobilization of endogenous retroviruses by recombination during the earliest stages of a germ-line invasion. *Proceedings of the National Academy of Sciences of the United States of America*, 115(34), 8609–8614. <https://doi.org/10.1073/pnas.1807598115>
- Mantere, T., Kersten, S., & Hoischen, A. (2019). Long-read sequencing emerging in medical genetics. *Frontiers in Genetics*, 10, e426. <https://doi.org/10.3389/fgene.2019.00426>
- Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., & Rothberg, J. M. (2005). Genome sequencing in open micro-fabricated high density picoliter reactors. *Nature*, 437(7057), 376–380. <https://doi.org/10.1038/nature03959>
- Maricic, T., Whitten, M., & Pääbo, S. (2010). Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS ONE*, 5(11), e14004. <https://doi.org/10.1371/journal.pone.0014004>
- Merker, J. D., Wenger, A. M., Sneddon, T., Grove, M., Zappala, Z., Fresard, L., Waggott, D., Utiramerur, S., Hou, Y., Smith, K. S., Montgomery, S. B., Wheeler, M., Buchan, J. G., Lambert, C. C., Eng, K. S., Hickey, L., Korfach, J., Ford, J., & Ashley, E. A. (2018). Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, 20(1), 159–163. <https://doi.org/10.1038/gim.2017.86>
- Meyer, M., & Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, 2010(6), pdb.prot5448. <https://doi.org/10.1101/pdb.prot5448>
- Nowrouzi, A., Dittrich, M., Klanke, C., Heinkelein, M., Rammling, M., Dandekar, T., & Rethwilm, A. (2006). Genome-wide mapping of foamy virus vector integrations into a human cell line. *The Journal of General Virology*, 87(Pt 5), 1339–1347. <https://doi.org/10.1099/vir.0.81554-0>
- Ochman, H., Gerber, A. S., & Hartl, D. L. (1988). Genetic applications of an inverse polymerase chain reaction. *Genetics*, 120(3), 621–623.
- Ondov, B. D., Bergman, N. H., & Phillippy, A. M. (2011). Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, 12, 385. <https://doi.org/10.1186/1471-2105-12-385>
- Ranzani, M., Cesana, D., Bartholomae, C. C., Sanvito, F., Pala, M., Benedicenti, F., Gallina, P., Sergi, L. S., Merella, S., Bulfone, A., Doglioni, C., von Kalle, C., Kim, Y. J., Schmidt, M., Tonon, G., Naldini, L., & Montini, E. (2013). Lentiviral vector-based insertional mutagenesis identifies genes associated with liver cancer. *Nature Methods*, 10(2), 155–161. <https://doi.org/10.1038/nmeth.2331>
- Renfree, M. B., Papenfuss, A. T., Deakin, J. E., Lindsay, J., Heider, T., Belov, K., Rens, W., Waters, P. D., Pharo, E. A., Shaw, G., Wong, E. S. W., Lefèvre, C. M., Nicholas, K. R., Kuroki, Y., Wakefield, M. J., Zenger, K. R., Wang, C., Ferguson-Smith, M., Nicholas, F. W., ... Worley, K. C. (2011). Genome sequence of an Australian kangaroo, *Macropus eugenii*, provides insight into the evolution of mammalian reproduction and development. *Genome Biology*, 12(8), R81. <https://doi.org/10.1186/gb-2011-12-8-r81>
- Rey-Iglesia, A., Gopalakrishnan, S., Carøe, C., Alquezar-Planas, D. E., Ahlmann Nielsen, A., Röder, T., Bruhn Pedersen, L., Næsberg-Nielsen, C., Sinding, M.-H. S., Fredensborg Rath, M., Li, Z., Petersen, B., Gilbert, M. T. P., Bunce, M., Mourier, T., & Hansen, A. J. (2019). MobiSeq: De novo SNP discovery in model and non-model species through sequencing the flanking region of transposable elements. *Molecular Ecology Resources*, 19, (2), 512–525. <https://doi.org/10.1111/1755-0998.12984>
- Schmidt, M., Schwarzwaelder, K., Bartholomae, C., Zaoui, K., Ball, C., Pilz, I., Braun, S., Glimm, H., & von Kalle, C. (2007). High-resolution insertion-site analysis by linear amplification-mediated PCR (LAM-PCR). *Nature Methods*, 4(12), 1051–1057. <https://doi.org/10.1038/nmeth.1103>
- Schmidt, M., Zickler, P., Hoffmann, G., Haas, S., Wissler, M., Muessig, A., & von Kalle, C. (2002). Polyclonal long-term repopulating stem cell clones in a primate model. *Blood*, 100(8), 2737–2743. <https://doi.org/10.1182/blood-2002-02-0407>
- Seeber, P. A., McEwen, G. K., Löber, U., Förster, D. W., East, M. L., Melzheimer, J., & Greenwood, A. D. (2019). Terrestrial mammal surveillance using hybridization capture of environmental DNA from African waterholes. *Molecular Ecology Resources*, 19(6), 1486–1496. <https://doi.org/10.1111/1755-0998.13069>
- Silver, J., & Keerikatte, V. (1989). Novel use of polymerase chain reaction to amplify cellular DNA adjacent to an integrated provirus. *Journal of Virology*, 63(5), 1924–1928. <https://doi.org/10.1128/JVI.63.5.1924-1928.1989>
- Symmons, O., & Spitz, F. (2013). From remote enhancers to gene regulation: charting the genome's regulatory landscapes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1620), 20120358. <https://doi.org/10.1098/rstb.2012.0358>
- Trowsdale, J., & Knight, J. C. (2013). Major Histocompatibility Complex Genomics and Human Disease. *Annual Review of Genomics and Human Genetics*, 14(1), 301–323. <https://doi.org/10.1146/annurev-genom-091212-153455>
- Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3 – New capabilities and interfaces. *Nucleic Acids Research*, 40(15), e115. <https://doi.org/10.1093/nar/gks596>
- Uren, A. G., Mikkers, H., Kool, J., van der Weyden, L., Lund, A. H., Wilson, C. H., Rance, R., Jonkers, J., van Lohuizen, M., Berns, A., & Adams, D. J. (2009). A high-throughput splinkerette-PCR method for the isolation and sequencing of retroviral insertion sites. *Nature Protocols*, 4(5), 789–798. <https://doi.org/10.1038/nprot.2009.64>
- Wang, M., Beck, C. R., English, A. C., Meng, Q., Buhay, C., Han, Y. I., Doddapaneni, H. V., Yu, F., Boerwinkle, E., Lupski, J. R., Muzny, D. M., & Gibbs, R. A. (2015). PacBio-LITS: A large-insert targeted sequencing method for characterization of human disease-associated chromosomal structural variations. *BMC Genomics*, 16, 214. <https://doi.org/10.1186/s12864-015-1370-2>

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Alquezar-Planas DE, Löber U, Cui P, Quedenau C, Chen W, Greenwood AD. DNA sonication inverse PCR for genome scale analysis of uncharacterized flanking sequences. *Methods Ecol Evol*. 2020;00:1–14. <https://doi.org/10.1111/2041-210X.13497>