nature research

Corresponding author(s): Rebecca Johnson

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.

For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Experimental design

1. Sample size

   Describe how sample size was determined.

   Samples of three koalas for genome sequencing were obtained by opportunistic collection and under appropriate animal ethics permits during routine veterinary care. Permits and protocols are detailed in Methods section, page 36, paragraph 1. Samples for population genomic aspect of analysis were chosen to cover the geographic distribution of koalas in Australia.

2. Data exclusions

   Describe any data exclusions.

   No data were excluded from analysis.

3. Replication

   Describe the measures taken to verify the reproducibility of the experimental findings.

   Replication is not used in this study. However there are a number of quality control measures presented in the paper (such as BUSCO analysis) that have been used to infer assembly and annotation quality in the koala genome presented in this work. See Supplementary file section 1.3 and supplementary Table 4. The PSMC analysis (figure 3)

4. Randomization

   Describe how samples/organisms/participants were allocated into experimental groups.

   Randomization was not relevant to this study. The work here presents a de novo genome assembly, annotation and associated analysis highlighting significant biological findings for the koala. The analysis was largely conducted on the genome of a single individual (however there are three animals sequenced in total, reported in Supplementary file section 1.2).

5. Blinding

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

   Blinding was not relevant to this study. which reports whole genome sequencing of three koala samples and associated analyses (which are focused on the single female koala 'Bilbo'. The population genomic analysis presented in this work included individually labeled samples during data collection but these did not influence the population genomic analysis.

Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ Test values indicating whether an effect is present *Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.*
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☒	☐ Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

All software (including commercially available programs) and parameters used in this work are detailed in the supplementary file and methods section of the main manuscript. Custom scripts are all publicly available at github (https://github.com/DrRebeccaJ/KoalaGenome)

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.

No unique materials were used. Koala samples used in the genome sequences presented in this work can be obtained from the Australian Museum frozen tissue collection and University of Sunshine Coast researchers.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

This is described in Supplementary Data section 2.2 paragraph 1. Centromeric antibodies (CENP-A and CREST) were used on koala (species Phascolarctos cinereus)

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

No eukaryotic cell lines were used

b. Describe the method of cell line authentication used.

No eukaryotic cell lines were used

c. Report whether the cell lines were tested for mycoplasma contamination.

No eukaryotic cell lines were used

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

No commonly misidentified cell lines were used

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.

Samples were opportunistically obtained during routine veterinary care. See Methods section page 36 and Supplementary data section 1.1 (paragraphs 1-2) for statement and animal care and ethics approval numbers

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

This study did not involve human research participants.

nature research

Corresponding author(s): Rebecca Johnson

ChIP-seq Reporting Summary

Form fields will expand as needed. Please do not leave fields blank.

Data deposition

1. For all ChIP-seq data:

☒ a. Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ b. Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

2. Provide all relevant data deposition access links.
The entry may remain private before publication.

Bioproject: PRJNA415832 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA415832)
And GEO submission: GSE111153 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE111153)

3. Provide a list of all files available in the database submission.

CENPA_R1.fastq.gz
CENPA_R2.fastq.gz
CREST_R1.fastq.gz
CREST_R2.fastq.gz
INPUT_R1.fastq.gz
INPUT_R2.fastq.gz
CENPA_bt2_vs_phaCin_unsw_v4.1_kd2_peaks.broadPeak.bed
CREST_bt2_vs_phaCin_unsw_v4.1_kd2_peaks.broadPeak.bed
md5sums.txt
Geo_Submission_Koala.xlsx

4. Provide a link to an anonymized genome browser session (e.g. UCSC), if available.

N/A (genome not on UCSC)

Methodological details

5. Describe the experimental replicates.

Two ChIP-seq experiments: 1 IP with CENP-A and 1 IP with CREST, each compared to input (replicates are antibodies, only limited quantity of 1 sample available)

6. Describe the sequencing depth for each experiment.

CENP-A paired end reads (2X75) 21,957,887; 98.51% mapped
CREST paired end reads (2X75) 33,421,909; 98.54% mapped
Input paired end reads (2X75) 39,171,541; 98.88% mapped

7. Describe the antibodies used for the ChIP-seq experiments.

CENP-A: Macropus eugenii anti-CenpA (rabbit) (derived by Biosynthesis), lot: AB1035-1
CREST: human anti-centromere protein IgG (Antibodies incorporated), cat# 15-235, lot: 441.20BK.82

8. Describe the peak calling parameters.

Reads were adapter clipped and trimmed with Trimmomatic 0.36 PE. Surviving paired reads were mapped to the koala genome (phaCin_unsw_v4.1) with bowtie 2 using the "very sensitive, paired end" parameters. Peaks were called by MACS2 (version 2.0.10.20131216) using the following parameters: –broad –keep-dup 2 –B –q 0.01

9. Describe the methods used to ensure data quality.

To identify candidate centromeric segments, two sliding window analyses were performed with a window size of 200kb and 20kb and a step size of 100kb and 10kb respectively. These regions were clustered using the heatmap.2 function in R (v3.2.5) package gplots (v3.0.1) with a high density of ChIP-seq peaks from CENP-A and CREST were identified through a manual curation of clusters of regions with similar peak densities. The repeats of these identified candidate regions were analyzed for biases between regions of interest and the remainder of the genome at the species, family and class level using the RepeatMasker reported metrics:

divergence from the model, total fraction of the bases in regions, frequency of repeat in regions and completeness of repeat (see below).

10. Describe the software used to collect and analyze the ChIP-seq data.

Surviving paired read files were converted to fasta files. The fasta files were broken up into smaller files of 1 million reads each. A single 1 million read fasta file from each pair was repeat masked with RepeatMasker 4.0.3 using the marsupial database as well as a koala denovo database. Repeat class and type was summarized using the buildSummary.pl script in the RepeatMasker utility script folder. The number of reads for each repeat in the RepeatMasker output file was normalized to the total number of repeats detected by RepeatMasker to obtain a frequency of detection for each repeat type.
Repeat content of the centromeric regions in phaCin_unsw_v4.1 was determined using RepBase annotated marsupial repeats and output from RepeatModeller analysis of phaCin_unsw_v4.1. The genome was masked using RepeatMasker to identify the location of repeats. To identify candidate centromeric segments, two sliding window analyses were performed with a window size of 200kb and 20kb and a step size of 100kb and 10kb respectively. These regions were clustered using the heatmap.2 function in R (v3.2.5) package gplots (v3.0.1) with a high density of ChIP-seq peaks from CENP-A and CREST were identified through a manual curation of clusters of regions with similar peak densities. The repeats of these identified candidate regions were analyzed for biases between regions of interest and the remainder of the genome at the species, family and class level using the RepeatMasker reported metrics: divergence from the model, total fraction of the bases in regions, frequency of repeat in regions and completeness of repeat. The RepeatMasker output file was converted into a bed file for use with bedtools (v2.25.0) and each candidate region was compared against the remaining candidates and the full genome as the background to identify the region with centromeric characteristic repeats compared to the background regions. To compare the candidate regions against the background regions, the Kolmogorov–Smirnov test (as implemented using ks.test from R) and Anderson-Darling test (as implemented in the kSamples package v1.2-4 from R) were used on each of the reported RepeatMasker metrics to identify which repeats were significantly different between the foreground and background regions. Using the average divergence from the repeat models and number of bases belonging to each repeat, the similarity of candidate region was visualized using multidimensional scaling and clustered using heatmaps using ggbiplot (v0.55) and gplots (v3.0.1) respectively.