Adaptation and conservation insights from the koala genome

Rebecca N. Johnson [1,2,30,31]*, Denis O'Meally [2,3,30], Zhiliang Chen [4,30], Graham J. Etherington [5], Simon Y. W. Ho [2], Will J. Nash [5], Catherine E. Grueber [2,6], Yuanyuan Cheng [2,7], Camilla M. Whittington [8], Siobhan Dennison [1], Emma Peel [2], Wilfried Haerty [5], Rachel J. O'Neill [9], Don Colgan [1], Tonia L. Russell [10], David E. Alquezar-Planas [1], Val Attenbrow [1], Jason G. Bragg [11,12], Parice A. Brandies [2], Amanda Yoon-Yee Chong [5,13], Janine E. Deakin [14], Federica Di Palma [5,15], Zachary Duda [9], Mark D. B. Eldridge [1], Kyle M. Ewart [1], Carolyn J. Hogg [2], Greta J. Frankham [1], Arthur Georges [14], Amber K. Gillett [16], Merran Govendir [8], Alex D. Greenwood [17,18], Takashi Hayakawa [19,20], Kristofer M. Helgen [1,21], Matthew Hobbs [1], Clare E. Holleley [22], Thomas N. Heider [9], Elizabeth A. Jones [8], Andrew King [1], Danielle Madden [3], Jennifer A. Marshall Graves [11,14,23], Katrina M. Morris [24], Linda E. Neaves [1,25], Hardip R. Patel [26], Adam Polkinghorne [3], Marilyn B. Renfree [27], Charles Robin [27], Ryan Salinas [4], Kyriakos Tsangaras [28], Paul D. Waters [4], Shafagh A. Waters [4], Belinda Wright [1,2], Marc R. Wilkins [4,10,30], Peter Timms [29,30] and Katherine Belov [2,30,31]

The koala, the only extant species of the marsupial family Phascolarctidae, is classified as 'vulnerable' due to habitat loss and widespread disease. We sequenced the koala genome, producing a complete and contiguous marsupial reference genome, including centromeres. We reveal that the koala's ability to detoxify eucalypt foliage may be due to expansions within a cytochrome P450 gene family, and its ability to smell, taste and moderate ingestion of plant secondary metabolites may be due to expansions in the vomeronasal and taste receptors. We characterized novel lactation proteins that protect young in the pouch and annotated immune genes important for response to chlamydial disease. Historical demography showed a substantial population crash coincident with the decline of Australian megafauna, while contemporary populations had biogeographic boundaries and increased inbreeding in populations affected by historic translocations. We identified genetically diverse populations that require habitat corridors and instituting of translocation programs to aid the koala's survival in the wild.

The koala is an iconic Australian marsupial, instantly recognizable by its round, humanoid face and distinctive body shape. Fossil evidence identifies as many as 15–20 species, following the divergence of koalas (Phascolarctidae) from terrestrial wombats (Vombatidae) 30–40 million years ago[1,2] (Supplementary Fig. 1). The modern koala, *Phascolarctos cinereus*, which first appeared in the fossil record ~350,000 years ago, is the only extant species of the Phascolarctidae. Like other marsupials, koalas give birth to

[1]Australian Museum Research Institute, Australian Museum, Sydney, New South Wales, Australia. [2]School of Life and Environmental Sciences, Faculty of Science, University of Sydney, Sydney, New South Wales, Australia. [3]Animal Research Centre, Faculty of Science, Health, Education & Engineering, University of the Sunshine Coast, Maroochydore, Queensland, Australia. [4]School of Biotechnology and Biomolecular Sciences, University of New South Wales, Kensington, New South Wales, Australia. [5]Earlham Institute, Norwich Research Park, Norwich, UK. [6]San Diego Zoo Global, San Diego, CA, USA. [7]UQ Genomics Initiative, University of Queensland, St Lucia, Queensland, Australia. [8]Sydney School of Veterinary Science, Faculty of Science, University of Sydney, Sydney, New South Wales, Australia. [9]Department of Molecular and Cell Biology and Institute for Systems Genomics, University of Connecticut, Storrs, CT, USA. [10]Ramaciotti Centre for Genomics, University of New South Wales, Kensington, New South Wales, Australia. [11]Research School of Biology, Australian National University, Canberra, Australian Capital Territory, Australia. [12]National Herbarium of New South Wales, Royal Botanic Gardens & Domain Trust, Sydney, New South Wales, Australia. [13]Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [14]Institute for Applied Ecology, University of Canberra, Bruce, Australian Capital Territory, Australia. [15]Department of Biological Sciences, University of East Anglia, Norwich, UK. [16]Australia Zoo Wildlife Hospital, Beerwah, Queensland, Australia. [17]Department of Wildlife Diseases, Leibniz Institute for Zoo and Wildlife Research, Berlin, Germany. [18]Department of Veterinary Medicine, Freie Universität Berlin, Berlin, Germany. [19]Department of Wildlife Science (Nagoya Railroad Co., Ltd.), Primate Research Institute, Kyoto University, Inuyama, Japan. [20]Japan Monkey Centre, Inuyama, Japan. [21]School of Biological Sciences, Environment Institute, Centre for Applied Conservation Science, and ARC Centre of Excellence for Australian Biodiversity and Heritage, University of Adelaide, Adelaide, South Australia, Australia. [22]Australian National Wildlife Collection, National Research Collections Australia, CSIRO, Canberra, Australian Capital Territory, Australia. [23]School of Life Sciences, La Trobe University, Bundoora, Victoria, Australia. [24]The Roslin Institute and R(D)SVS, University of Edinburgh, Easter Bush, Midlothian, UK. [25]Royal Botanic Garden Edinburgh, Edinburgh, UK. [26]John Curtin School of Medical Research, Australian National University, Acton, Australian Capital Territory, Australia. [27]School of BioSciences, University of Melbourne, Melbourne, Victoria, Australia. [28]Department of Translational Genetics, The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus. [29]Faculty of Science, Health, Education & Engineering, University of the Sunshine Coast, Maroochydore, Queensland, Australia. [30]These authors contributed equally: Rebecca N. Johnson, Denis O'Meally, Zhiliang Chen, Marc R. Wilkins, Peter Timms, Katherine Belov. [31]These authors jointly supervised this work: Rebecca N. Johnson, Katherine Belov. *e-mail: rebecca.johnson@austmus.gov.au

underdeveloped young. Birth occurs after just 35 d of gestation, with young lacking immune tissues or organs. Their immune system develops while they are in the pouch, meaning survival during early life depends on immunological protection provided by mothers' milk.

A specialist arboreal folivore feeding almost exclusively from *Eucalyptus* spp., the koala has a diet that would be toxic or fatal to most other mammals[3]. Due to the low caloric content of this diet, the koala rests and sleeps up to 22 h a day[4]. A detailed understanding of the mechanisms by which koalas detoxify eucalyptus and protect their young in the pouch has been elusive, as there are no koala research colonies and access to milk and tissue samples is opportunistic. The genome enables unprecedented insights into the unique biology of the koala, without having to harm or disturb an animal of conservation concern.

The genome also enables a holistic, scientifically grounded approach to koala conservation. Australia has the highest mammal extinction record of any country during the Anthropocene[5], and koala numbers have plummeted in northern parts of its range since European settlement of the continent[6], but increased in southern sections of the range, notably in parts of Victoria and South Australia. The uneven response of koala populations throughout its range is one of the most difficult issues in its management[7]. The species was heavily exploited by a pelt trade (1870s to late 1920s), which harvested millions of animals[6,8,9]. Today, the threats are primarily due to loss and fragmentation of habitat, urbanization, climate change and disease. Current estimates put the number of koalas in Australia at only 329,000 (range 144,000–605,000), and a continuing decline is predicted[6]. Koalas present a complex conservation conundrum: in the north, causes of decline include ongoing habitat fragmentation, urbanization and disease. However, decline in the south has followed a different path[10], with widespread, often sequential, translocations (1920–1990s) from a limited founder population, which has resulted in genetically bottlenecked populations that are overabundant to the point of starvation in some areas[11]. There are marked differences in the degree to which threats affect each population, thereby cautioning against one prescription for population recovery.

Adding to the complexity of koala conservation is the impact of disease, specifically koala retrovirus (KoRV) and *Chlamydia*. KoRV is thought to have arrived in Australia via a putative murine vector before cross-species transmission[12,13]. It is now prevalent in northern koalas and appears to be spreading to southern populations[14]. Some strains appear to be more virulent than others and are putatively associated with an increase in neoplastic disease[15]. Similarly, *Chlamydia*, which in some individuals causes severe symptoms yet in others remains asymptomatic, may have crossed the species barrier from introduced hosts such as domestic sheep and cattle following European settlement[16]. A complete koala genome offers insights into the species' genetic susceptibility to these diseases, provides the genomic basis for innovative vaccines, and can underpin new conservation management solutions that incorporate the species' population and genetic structure, such as facilitating gene flow via habitat connectivity or translocations.

Results

Genome landscape. Koalas have 16 chromosomes, differing from the ancestral marsupial $2n = 14$ karyotype by a simple fission of ancestral chromosome 2 giving rise to koala chromosomes 4 and 7[17]. We sequenced the complete genome using 57.3-fold PacBio long-read coverage, generating a 3.42 Gb reference assembly. The primary contigs from the FALCON assembly (representing homozygous regions of the genome) yielded genome version phaCin_unsw_v4.1. This comprised 3.19 Gb, including 1,906 contigs with an N50 of 11.6 Mb and the longest at 40.6 Mb. The heterozygous regions of the genome (representing the alternate contigs

from the assembly) totaled 230 Mb, with an N50 of 48.8 kb (Table 1, Supplementary Tables 1–3 and Methods). Approximately 30-fold coverage of Illumina short reads was used to polish the assembly. BioNano optical maps plus additional conserved synteny information for marsupials were used for scaffolding[18] to assemble long-read contigs into 'virtual' chromosome scaffolds ('super-contigs') (Supplementary Tables 4 and 5 and Supplementary Note). The largest super-contig spanned approximately half of koala chromosome 7 (Supplementary Fig. 2).

Our long-read-based sequence presented the opportunity to identify and study centromeres, which are multi-megabase regions that are challenging to construct in eutherian (for example, human and mouse)[19] genome assemblies due to intractable higher order arrays of satellites. Centromeres are smaller in marsupials than in eutherians, and as such are more amenable to analysis[20]. Chromatin immunoprecipitation and sequencing using antibodies to centromeric proteins (CENP-A and CREST)[21] enabled the identification of scaffolds containing putative centromeric regions (Supplementary Fig. 3) and the characterization of known and new repeats, including composite elements within koala centromeric domains (Supplementary Table 6–10) that lack the previously annotated retroelement, kangaroo endogenous retrovirus (KERV), found in some tammar wallaby centromeres[22]. Koala centromeres span a total of 2.6 Mb of the koala haploid genome, equivalent to an average of 300 kb of centromeric material per chromosome. Like those of other species with small centromeres[19,20,23,24], koala centromeres lack higher order satellite arrays (Supplementary Tables 7–10). Among the newly identified repeats, some are similar to composite elements recently described in gibbon centromeres[25], where absence of higher order satellite arrays accompanied the evolution of new composite elements with putative centromere function. The composition of the koala centromere therefore supports mounting evidence that transposable elements represent a major, functional component of small centromeres when higher order satellite arrays are absent[20,24,25].

Interspersed repeats account for approximately 47.5% of the koala genome; 44% of these are transposable elements (Supplementary Table 11). As in other mammalian genomes, short interspersed nuclear elements (SINEs) and long interspersed nuclear elements (LINEs) are the most numerous elements (35.2% and 28.9% of total number of elements, respectively), with LINEs making up 32.1% of the koala genome. The long-read sequence assembly also enabled full characterization and annotation of repeat-rich long noncoding RNAs, including *RSX*, which mediates X chromosome inactivation in female marsupials[26]. Koala *RSX* represents the first marsupial *RSX* to be fully annotated and to have its structure predicted (Supplementary Fig. 4 and Supplementary Note). As expected, it was expressed in all female tissues, but in no male tissues[27].

The assembled koala genome has very high coverage of coding regions: we recovered 95.1% of 4,104 mammalian benchmarking universal single-copy orthologs (BUSCOs)[28], the highest value for any published marsupial genome (Supplementary Table 5) and comparable with that of the human assembly (GRCh38, which scores 94.1% of orthologs). Analysis of gene family evolution using a maximum-likelihood framework identified 6,124 protein-coding genes in 2,118 gene families with at least two members in koala. Among these, 1,089 have more gene members in koala than in any of the other species (human, mouse, dog, tammar wallaby, Tasmanian devil, gray short-tailed opossum, platypus, chicken; Supplementary Fig. 5).

Having characterized the genome, we undertook detailed analyses of key genes and gene families to gain insights into the genomic basis of the koala's highly specialized biology. Gene families of particular interest were those that encode proteins involved in induced ovulation, those proteins involved in the complex lactation process, those proteins responsible for immunity, and those enzymes that enable the koala to subsist on a toxic diet.

Table 1 | Comparison of assembly quality between koala genome assembly phaCin_unsw_v4.1 and published marsupial and monotreme genomes

Species	Genome size (Gb)	G+C content (%)	No. scaffolds	Scaffold N50 (kb)	Reference
Koala phaCin_unsw_v4.1 (female Bilbo)	3.42	39.0	1,906[a] 5,525[b] (contigs)	11,589 (contig)	This study
Platypus (*Ornithorhynchus anatinus*)	2.3	45.5	200,283	959	Warren et al. 2008[82]
Gray short-tailed opossum (*Monodelphis domestica*)	3.48	37.7	5,223	59,810	Mikkelsen et al. 2007[83]
Tammar wallaby (*Notamacropus eugenii*)	2.7	38.8	277,711	37	Renfree et al. 2011[84]
Tasmanian devil (*Sarcophilus harrisii*)	3.17	36.4	35,974	1,847	Murchison et al. 2012[85]

[a]Homozygous. [b]Heterozygous.

Ability to tolerate a highly toxic diet. The koala's diet of eucalyptus leaves contains high levels of plant secondary metabolites[29], phenolic compounds[30] and terpenes (for example, ref. [31]) that would be lethal to most other mammals[32]. Koalas thus experience little competition for food resources. *Eucalyptus grandis* shows substantial expansion in terpene synthase genes relative to other plant genomes[33]. Eucalypt toxicity is therefore likely to have exerted selection pressure on the koala's ability to metabolize such xenobiotics, so we searched for genes encoding enzymes with a detoxification function and investigated sequence evolution at these loci.

Cytochrome P450 monooxygenase (*CYP*) genes represent a multi-gene superfamily of heme-thiolate enzymes that play a role in detoxification through phase 1 oxidative metabolism of a range of compounds including xenobiotics[34]. These genes have been identified throughout the tree of life, including in plants, animals, fungi, bacteria and viruses[35]. In the koala genome we found two lineage-specific monophyletic expansions of the cytochrome P450 family 2 subfamily C (*CYP2C*s, 31 members in koala) (Fig. 1a). The functional importance of these *CYP2C* genes was further demonstrated through analysis of expression in 15 koala transcriptomes from two koalas, showing particularly high expression in the liver, consistent with a role in detoxification (Supplementary Fig. 6).

Comparing *CYP2C* gene context in mouse versus koala identified conserved flanking markers strongly suggestive of tandem duplication (Fig. 1b). Further sequence-level analysis of the *CYP* expansions indicated that most conserved regions are under strong purifying selection (Fig. 1c). However, there is evidence that individual *CYP* codons have experienced episodic diversifying selection while purifying selection shapes the rest of the gene (Fig. 1c–h, Supplementary Note and Supplementary Tables 12 and 13). Adaptive expansion of *CYP2C* and maintenance of duplicates appear to have worked in concert, resulting in higher enzyme levels for detoxification while the interplay between purifying and diversifying selection resulted in neofunctionalization within the *CYP*s. Such adaptations enable koalas to detoxify their highly specialized diet rich in plant secondary metabolites.

The characterization of koala *CYP2C*s has significant therapeutic potential. The high expression levels of *CYP2C* genes in the liver helps to explain why meloxicam, a nonsteroidal anti-inflammatory drug (NSAID) known to be metabolized by the protein product of *CYP2C* in humans[36,37] and frequently used for pain relief in veterinary care, is so rapidly metabolized in the koala and a handful of other eucalypt-eating marsupials (common brushtail possum and eastern ringtail possum) compared with eutherian species[37,38]. It is expected that other NSAIDs are also rapidly metabolized in koalas and have little efficacy at suggested doses[39]. Anti-chlamydia antibiotics such as chloramphenicol are degraded rapidly by koalas;

treatment with a single dose applicable to humans is insufficient in koalas, which require a daily dose for up to 30 to 45 d. This discovery of *CYP2C* gene expression levels will inform new research into the pharmacokinetics of medicines in koalas.

Taste, smell and food choice. Like many specialist folivores, koalas are notoriously selective feeders, making food choices both to target nutrients and to avoid plant secondary metabolites[40]. Koalas have been observed to sniff leaves before tasting them[41], and their acute discrimination has been correlated with the complexity and concentration of plant secondary metabolites[42]. This suggests an important role for olfaction and vomerolfaction, as well as taste. While most herbivores circumvent plant chemical defenses by detoxifying one or a few compounds[43], the complexity of eucalyptus plant secondary metabolites, in combination with the terpene expansion in eucalypts, led us to hypothesize that the koala requires enhanced capabilities both in specialist detection and in plant secondary metabolite detoxification. We therefore investigated the genomic basis of the koala's taste and smell senses, finding multiple gene family expansions that could enhance its ability to make food choices.

We report an expansion of one lineage of vomeronasal receptor type 1 (*V1R*) genes associated with the detection of nonvolatile odorants (Supplementary Note). There are six such genes in koala, compared with only one in the Tasmanian devil and gray short-tailed opossum, and none found in tammar wallaby, human, mouse, dog, platypus or chicken. The expansion of one lineage of *V1R* genes is consistent with the koala's ability to discriminate among diverse plant secondary metabolites.

Surprisingly, given the degree of its dietary specialization, the olfactory receptor genes ($n = 1,169$) characterized in koala had a gene repertoire that was slightly smaller than that of gray short-tailed opossum (1,431 genes), tammar wallaby (1,660 genes) and Tasmanian devil (1,279 genes) (Supplementary Note). This may be understood in the context of relaxed selection on olfactory receptors among dietary specialists[44].

We also report genomic evidence of expansions within the taste receptor families that would enable the koala to optimize ingestion of leaves with a higher moisture and nutrient content in concert with the concentration of toxic plant secondary metabolites in their food plants. The koala's ability to 'taste water' is potentially enhanced by an apparent functional duplication of the aquaporin 5 gene[45–47] (Supplementary Table 14 and Supplementary Note).

The *TAS2R* family has a role in 'bitter' taste, enabling recognition of structural toxins such as terpenes, phenols and glycosides. These are found in various levels in eucalypts as plant secondary metabolites[3,30,31,48]. In marsupials, the *TAS2R* family includes the orthologous repertoires from eutherians, as well as three specific
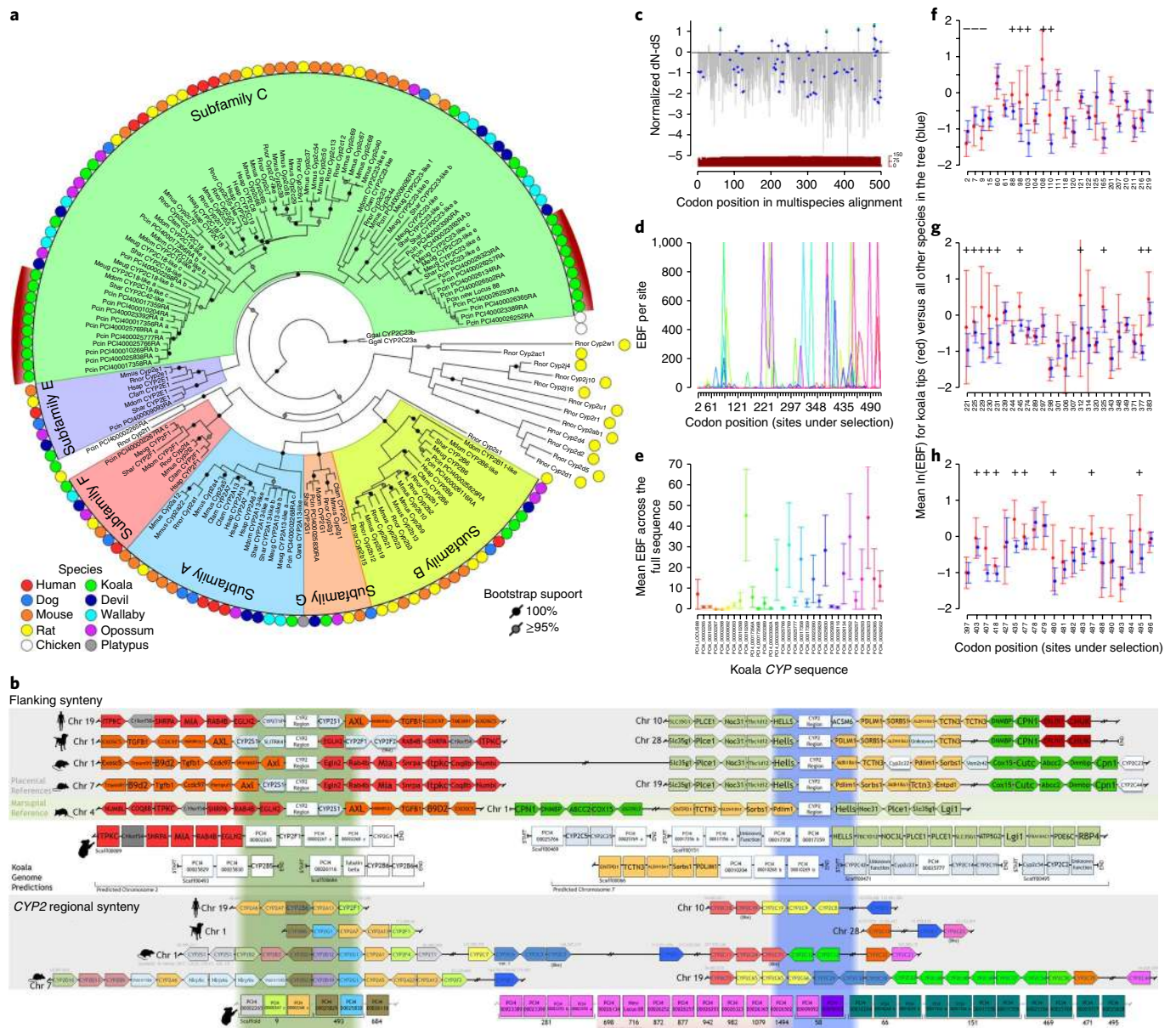
Fig. 1 | Analysis of cytochrome P450 family 2 subfamily C gene family. a, Phylogenetic tree of *CYP2* gene family in koala (Pcin; 31 *CYP2* members), compared with marsupials: tammar wallaby (Meug), Tasmanian devil (Shar), gray short-tailed opossum (Mdom); eutherian mammals: human (Hsap), rat (Rnor), mouse (Mmus), dog (Cfam); monotreme: platypus (Oana); and outgroup chicken (Ggal). Two independent monophyletic expansions are seen in koala in the *CYP2C* subfamily (highlighted by red arcs). **b**, *CYP* synteny map showing expansion of *CYP2C* genes in koala as compared to eutherians (human, dog, rat, mouse) and another marsupial (opossum). **c**–**h**, Selection analysis of *CYP* gene expansion. **c**, Normalized d*N*-d*S* (SLAC (single-likelihood ancestor counting) method) across the alignment of 152 *CYP* sequences (only sites with data in koala and at least one other species; red bars show sequence depth). Points indicate statistically significant (threshold α = 0.1) evidence for codons under selection. Four sites show positive selection across entire tree (SLAC; green points); 70 sites show episodic selection (MEME (mixed effects model of evolution); blue diamonds). **d**, Comparison of episodic selection on particular codons across koala *CYP* genes (*n* = 31); *x* axis shows codons with evidence of statistically significant selection anywhere on the tree (identified in **c**). **e**, Comparison of mean episodic selection among koala *CYP* genes (*n* = 70). Points indicate mean empirical Bayes factor (EBF) for sites under selection for each sequence; error bars, 95% confidence interval. **f**–**h**, Mean EBF (natural log transformed, EBF values of 0 excluded) for koala tree tips (*n* = 31; red) relative to all others (*n* = 121 in 9 species (see Methods), blue). Points show mean, error bars ± 95% confidence interval, evaluated as 1.96 × s.e.m. (using sequence depth as sample size; red bars in **c**). Codon positions on *x* axis refer to multispecies alignment from **c**. Symbols above each point indicate that the mean value for koala site falls outside the 95% confidence interval for all other species (above, "+"; below, "−"; two-tailed test at α = 0.05). Raw statistics shown (unadjusted for multiple comparisons).

expansions in the last common ancestor shared by all marsupials[49,50] (Fig. 2). Large koala-specific duplications in four marsupial orthologous groups have produced a large koala *TAS2R* repertoire of 24 genes (Fig. 2). The koala has more *TAS2Rs* than any other Australian marsupial, and among the most of all mammal species[49,50], including paralogs of human and mouse receptors whose agonists are toxic glycosides (Supplementary Table 15 and Supplementary Note). The *TAS1R* gene families, responsible for sweet taste and umami amino
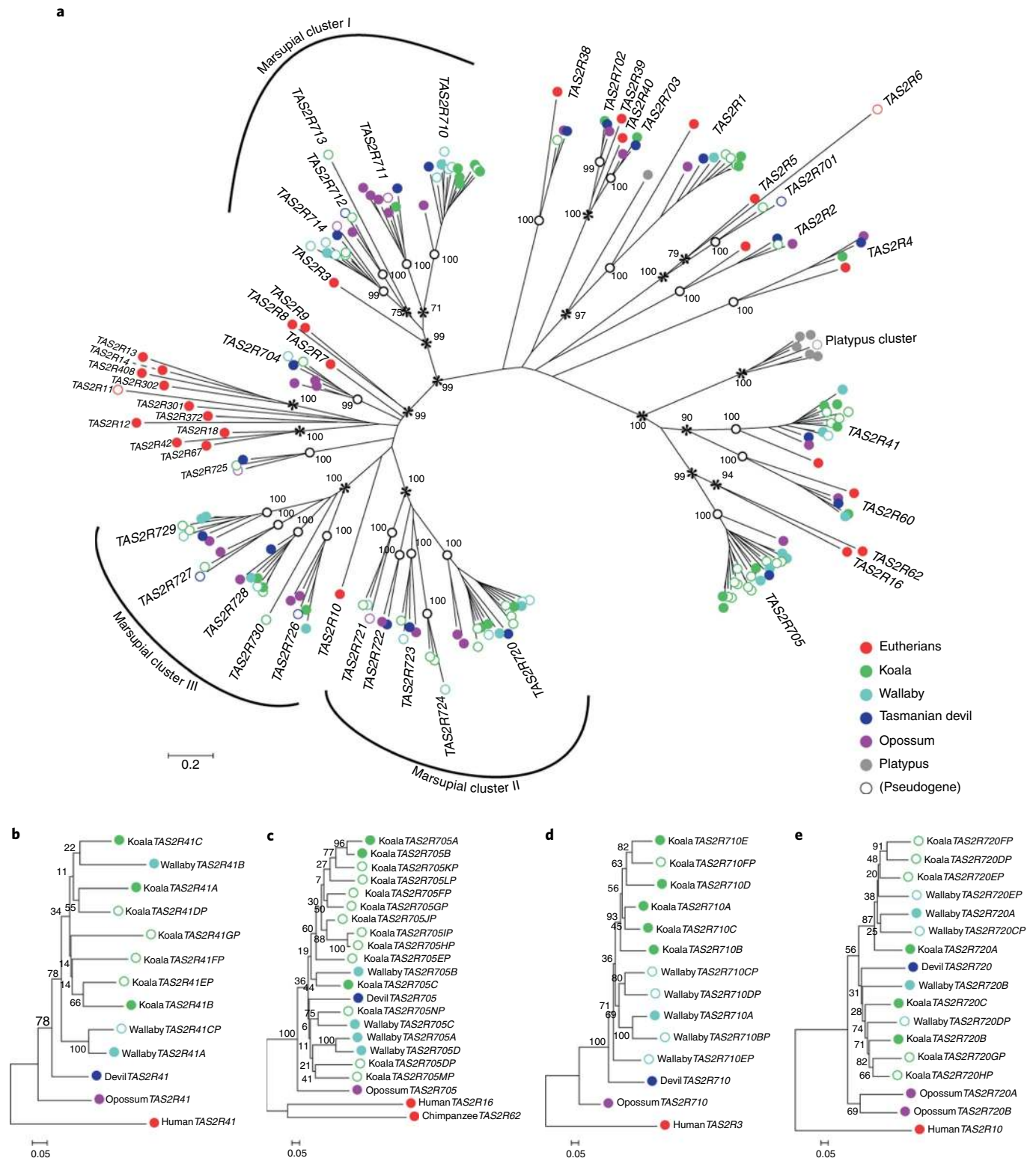
Fig. 2 | Taste receptor analysis in koalas and other mammals identifies three marsupial-specific expansions and further koala-specific duplications.
TAS2R genes are responsible for bitter taste perception. a, Maximum-likelihood tree of *TAS2R*s (including pseudogenes) in the four marsupials, where the sequences contained 250 amino acids. 28 representative *TAS2R*s of orthologous gene groups (OGGs) in eutherians (red circles) and 7 platypus *TAS2R*s (gray circles) were also used. There were 27 distinct marsupial OGGs (supported by ≥99% bootstrap values), where the nodes of OGG clades are indicated by white open circles. Bootstrap values of ≥70% in the nodes connecting OGG clades are indicated by asterisks. There are three marsupial-specific clusters (I, II and III) where massive expansion events occurred in the common ancestor of marsupials after their split from eutherian ancestors. b–e, Reconstructed maximum-likelihood trees of *TAS2R* orthologs in which there are more than two duplicates of koala *TAS2R*s: b, *TAS2R41*; c, *TAS2R705*; d, *TAS2R710*; and e, *TAS2R720*. Genomic structures of the umami and sweet taste receptor *TAS1R*s were also analyzed and found to be functional in koala (see Supplementary Note).

acid perception, have previously been reported as pseudogenized in eutherians with highly specialized diets, such as the giant panda[51]. In the koala, however, we found that all *TAS1R* genes are putatively functional (Supplementary Fig. 7).

Genomics of an induced ovulator. Koala reproduction is of particular interest because the koala is an induced ovulator[52], with key genes controlling female ovulation (*LHB*, *FSHB*, *ERR1*, *ERR2*), as well as prostaglandin synthesis genes important in parturition and ejaculation (*PTGS1*, *PTGS2*, *PTGS3*) (Supplementary Note). We identified genes putatively involved in the induction of ovulation in the female by male seminal plasma (*NGF*), and in coagulation of seminal fluid (*ODC1*, *SAT1*, *SAT2*, *SMOX*, *SRM*, *SMS*) (Supplementary Note), which may function to prevent sperm leakage from the female reproductive tract in this arboreal species.

Genomic characterization of koala milk. A koala young is about the size of a kidney bean and weighs < 0.5 g. It crawls into the mother's posteriorly opening pouch and attaches to a teat, where it remains for 6–7 months. It continues to suck after it has left the pouch until about a year old.

Analysis of the genome, in conjunction with a mammary transcriptome and a milk proteome, enabled us to characterize the main components of koala milk (Supplementary Fig. 8, Supplementary Table 16, Supplementary Note and ref. [53]). The high-quality assembly of the genome allowed both the identification of marsupial-specific genes and determination of their evolutionary origins based on their genomic locations. For instance, we found that there are four Late Lactation Protein (*LLP*) genes tightly linked to both trichosurin and β-lactoglobulin (Supplementary Fig. 8), potentially allowing marsupials to fine-tune milk protein composition across the stages of lactation to meet the changing needs of their young. Additionally, the koala marsupial milk 1 (*MM1*) gene, a novel marsupial gene, is located close to the gene encoding very early lactation protein (*VELP*), an ortholog of *Glycam1* (or *PP3*) that encodes a eutherian antimicrobial protein[53] (Supplementary Fig. 8). In eutherians, this region contains an array of short glycoproteins that have antimicrobial properties and are found in secretions such as milk, tears and sweat. We propose that *MM1* has an antimicrobial role in marsupial milk, along with three other short novel genes located in the same region. We also detected expansions in another antimicrobial gene family, the cathelicidins.

Koala immunome and disease. At the time of European settlement, koalas were widespread in eastern mainland Australia, from north Queensland to the southeastern corner of South Australia. Today they are mainly confined to the east coast and are listed as 'vulnerable' under Australia's *Environment Protection and Biodiversity Conservation Act 1999*[54]. There is strong evidence to suggest that some fragmented populations of koalas are already facing extinction, particularly in formerly densely populated koala territories in southeast Queensland and northern New South Wales. A major challenge for the conservation of these declining koala populations is the high prevalence of disease, especially that caused by the obligate intracellular bacterial pathogen *Chlamydia pecorum*, which is found across the geographic range, with the exception of some offshore islands[55]. A main challenge for managing these populations has been the lack of knowledge about the koala immune response to disease. Recent modeling suggests the best way to stabilize heavily affected koala populations is to target disease[56].

The long-read-based genome enabled the de novo assembly of complex, highly duplicated immune gene families and comprehensive annotation of immune gene clusters[53,57,58]. These include the major histocompatibility complex (*MHC*)[59], as well as T cell receptors (*TCR*), immunoglobulin (*IG*) (Supplementary Fig. 9, Supplementary Tables 17 and 18, and Supplementary Note), natural killer cell (NK) receptor[58] and defensin[60] gene clusters. Together these findings provide a starting point for new disease research and allow us to interrogate the immune response to the most significant pathogen of the koala, *C. pecorum*.

Of the more than 1,000 koalas arriving annually at wildlife hospitals in Queensland and New South Wales, 40% have late-stage chlamydial disease and cannot be rehabilitated. Annotation of koala immune genes enabled us to study variation within candidate genes known to play a role in resistance and susceptibility to chlamydia infection in other species (Supplementary Tables 18–20). Preliminary case/control association tests for five koalas involved in a chlamydia vaccination trial showed that the MHCII *DMA* and *DMB* genes, as well as the *CD8-a* gene, may be involved in differential immune responses to chlamydia vaccine (Supplementary Table 21 and Supplementary Note). We also conducted differential expression analysis of RNA sequencing (RNA-seq) data from conjunctival tissue collected from koalas at necropsy, both with and without signs of ocular chlamydiosis, showing that in diseased animals, 1,508 of the 26,558 annotated genes (5.7%) were twofold upregulated, while 685 (2.6%) were downregulated by greater than twofold when compared with healthy animals (Supplementary Fig. 9 and Supplementary Note). In diseased animals, upregulated genes were associated with Gene Ontology (GO) terms for a range of immunological processes, including signatures of leukocyte infiltration (Supplementary Fig. 9). Immune responses in the affected conjunctivas were directed at T_H1 rather than T_H2 responses. Proinflammatory mediators such as *CCL20*, *IL1α*, *IL1β*, *IL6* and *SSA1* were also upregulated. As in human trachoma, this cascade of proinflammatory products may help to clear the infection but may also lead to tissue damage in the host[61]. Furthermore, resolution of human trachoma infection is thought to require a IFN-γ driven T_H1 response[62], and in diseased koalas we found that IFN-γ was upregulated 4.7-fold in the conjunctival tissue. These annotated koala immune genes will now help us to define features of protective versus pathogenic immunological responses to the disease and may be invaluable for effective vaccine design.

Koala genomes are undergoing genomic invasion by koala retrovirus (KoRV)[63], which is spreading from the north of the country to the south. Both endogenous (germline transmission) and exogenous (infectious 'horizontal' transmission) forms are extant[64]. Our results provide a comprehensive view of KoRV insertions in the koala genome. We found a total of 73 insertions in the phaCin_unsw_4.1 assembly (Supplementary Table 22). It is likely that most of these 73 loci are endogenous, consistent with our observation of integration breakpoint sequences that are shared with one or both of the other koala genomes reported (Supplementary Tables 23 and 24).

We investigated the sites of KoRV insertion to define their proximity to protein-coding genes and explore possible disruptions. This analysis identified insertions into 24 protein-coding genes (Supplementary Table 25). However, none is likely to disrupt protein-coding capacity, since 22 insertions are in introns and the other two are in 3′ untranslated regions. Transcription proceeding from the proviral long terminal repeat (LTR) could possibly affect the transcription of the host genes.

Understanding the genetics of host resistance to chlamydia and the etiology of the retrovirus will help inform the development of vaccines against both diseases, as well as translocation strategies.

Genome-informed conservation. Broad-scale population management of koalas is critical to conservation efforts. This is challenging because distribution models are not easily generalized across bioregions, and further complicated by the unique regional conservation
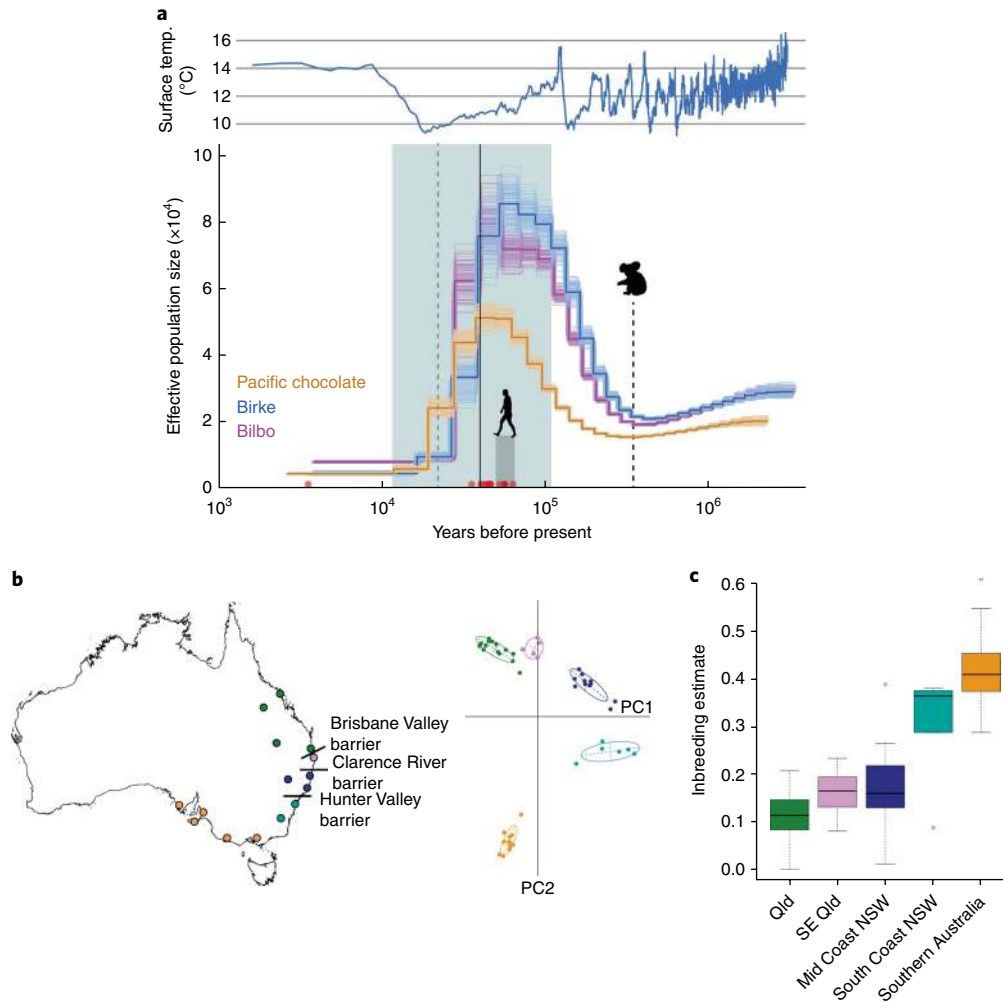
Fig. 3 | Analysis koala populations using genome-mapped markers. a, Top: plot of surface temperature (temp.) over past 3 million years based on a five-point running mean of δ^{18}O data[76]. Bottom: population demographic history inferred from diploid sequences of three koalas (females 'Pacific Chocolate' and 'Bilbo', male 'Birke') using the pairwise sequential Markovian coalescent (PSMC) method. Koala silhouette indicates earliest fossil record of modern koala[2]. Gray shading, human arrival in Australia[77] (see ref. [78]); red circles, estimated extinction times of 16 megafaunal genera in mainland Australia[79]; aqua area, last glacial period; vertical dashed green line, last glacial maximum; vertical solid black line, first koala population declines 40,000 years ago. Dark colored lines are estimated from genome data; lighter lines, plots inferred from 100 bootstrap replicates. A mutation rate of 1.45×10^{-8} mutations per site per generation and 7-year generation time were assumed. **b**, Right, principal component (PC) analysis (including 95% inertia ellipses) of 1,200 SNPs in 49 wild koalas from throughout Australia. Left, geographic clustering of wild koalas in eastern Australia in relation to proposed biogeographic barriers[68,72], highlighting known historic barriers to gene flow, the Brisbane and Clarence River Valleys, but also suggesting a role for the Hunter Valley. The cluster of genetically similar southern koalas reflects a recent history of widespread translocation[8]. **c**, Average inbreeding coefficient (*F*) (calculated by TrioML[80,81]) of 49 wild koalas. Qld, Queensland; SE, southeast; NSW, New South Wales. *P* values arising from linear modeling represent significant differences in mean *F* between regions (***$P < 0.001$; **$P < 0.01$). There is a high correlation between geographic distance and genetic distance (Mantel test: $r^2 = 0.4898$), indicating that genetic rescue between populations is feasible. Center lines, median; box limits, upper and lower quartiles. Upper whisker = min(max(*x*), Q_3 + 1.5 × IQR), lower whisker = max(min(*x*), Q_1 – 1.5 × IQR); i.e., upper whisker = upper quartile + 1.5 × box length, lower whisker = lower quartile – 1.5 × box length; circles, outliers. Linear modeling indicated that mean *F* differed significantly between several regions (Mid-coast NSW–Southern Australia, $P = 0.000524$; Qld–Southern NSW, $P = 0.00237$; Qld–Southern Australia, $P = 0.00000107$; SE Qld–Southern Australia, $P = 0.006596$).

issues described above. Since it is not possible to generalize management, it is imperative that decisions are informed by empirical data relevant to each bioregion.

Analysis of the koala genome provided the unique opportunity to combine historical evolutionary data with high-resolution contemporary population genomic markers to address these management challenges. To infer the ancient demographic history of the species, we analyzed the long-read reference genome and short-read data from two other koalas, using the pairwise sequentially Markovian coalescent (PSMC) method[65] (Fig. 3a, Supplementary Fig. 10 and

Methods). The data show that the modern koala, which appeared in the fossil record 350,000 years ago[2], underwent an initial increase in population, followed by a rapid and widespread decrease in population size ~30,000–40,000 years ago. This is consistent with fossil evidence of rapid declines in multiple Australian species, including the extinct megafauna, 40,000–50,000 years ago[66] and 30,000–40,000 years ago[67]. The koala was thus one of a number of species affected by decline during this time that did not ultimately become extinct[67].

Distinct PSMC profiles of the koalas from two geographic areas and their failure to coalesce suggests some regional differences in

koala populations, including impediments to gene flow (Fig. 3a). Regional differentiation was also detected in analyses of mtDNA[68,69], although over a shorter time scale.

We analyzed populations of recent koala samples using 1,200 SNPs derived from targeted capture libraries mapped to the koala genome (Supplementary Note). We found notable levels of genetic diversity with limited fine-scale differentiation consistent with long-term connectivity across regions. We found evidence of low genetic diversity in southern koalas, consistent with a recent history of sequential translocations[8,68,70,71] (Fig. 3b,c). At a continental scale, we show biogeographic barriers to gene flow associated with the Brisbane Valley and Clarence River, as identified by mtDNA studies[68,72], and find a barrier associated with the Hunter Valley, which was not previously known in koalas (Fig. 3b). Levels of inbreeding varied across regions (Fig. 3c), but the northern populations most under threat in New South Wales and Queensland show high levels of genetic diversity.

The information generated here provides a foundation for a conservation management strategy to maintain gene flow regionally while incorporating the genetic legacy of biogeographic barriers. Furthermore, the contrast in genome-wide levels of diversity between southern and northern populations highlights the detrimental consequences of the unmonitored use of small isolated populations as founders for reestablishing and/or rescuing of populations on genome-wide levels of genetic diversity. Low levels of genetic diversity in southern koalas have been associated with genetic abnormalities consistent with inbreeding depression, including testicular abnormalities[73].

Now that we understand the consequences of past translocations, and the existing genetic structure, it is clear that maintaining and facilitating gene flow via habitat connectivity will be the most effective means of ensuring genetically healthy koala populations over the long term. However, where more intensive measures such as translocation are required to rescue genetically depauperate southern populations, these tools and data provide the basis for decisions that maximize benefits while minimizing risks[74,75]. Future utility of these SNPs will also include tracking of individual pedigrees in captive koala populations and in those wild populations being intensively monitored.

The koala genome offers insights into historic and contemporary population dynamics, providing evolutionary and genetic context for a species that is the focus of considerable management actions and resources. By providing a deeper understanding of disease dynamics and population genetic processes, including the maintenance and monitoring of gene flow, this genomic information will enable the development of strategies necessary to preserve the species, from the preservation of habitat corridors through to the genetic rescue of isolated populations. As members of government advisory committees, some of the authors have initiated inclusion of genomic information into the New South Wales Koala Strategy. This will be used to inform koala management in the state with the goal of securing koalas in the wild for the future.

Discussion

The koala genome provides the highest quality marsupial genome to date. This assembly has enabled insights into the colonization of the koala genome by an exogenous retrovirus and revealed the architecture of the immune system, necessary to study and treat emerging diseases that threaten koala populations. A greater understanding of genetic diversity across the species will guide the selection of individuals from genetically healthy northern populations to augment genetically restricted populations in the south, bearing in mind that chlamydia has not been detected on some offshore islands, so risk assessment should be carried out before embarking on translocations. Sequencing the genome has advanced our understanding of the unique biology of the koala,

including detoxification pathways and innovations in taste and smell to enable food choices in an obligate folivore. Long-term survival of the species depends on understanding the impacts of disease and management of genetic diversity, as well as the koala's ability to source moisture and select suitable foraging trees. This is particularly important given the koala's narrow food range, which makes it especially vulnerable to a changing climate. The genome provides a springboard for conservation of this biologically unique and iconic Australian species.

URLs. FALCON assembly algorithm, https://github.com/PacificBiosciences/FALCON-integrate/; FALCON (v 0.3.0), http://falconframework.org/; RepeatMasker (v 4.0.3), http://www.repeatmasker.org/; RepeatModeler, http://www.repeatmasker.org/RepeatModeler/; RepBase (v 2015-08-07), http://www.girinst.org/repbase/; MAKER, http://www.yandell-lab.org/software/maker.html; Trinity (v 2.3.2), https://github.com/trinityrnaseq/trinityrnaseq/; SNAP, http://archive.broadinstitute.org/mpg/snap/; GeneMark, http://opal.biology.gatech.edu/GeneMark/; Augustus, http://bioinf.uni-greifswald.de/augustus/; NCBI Blast (v 2.3.0), https://blast.ncbi.nlm.nih.gov/Blast.cgi; OrthoMCL (v 2.0.9), http://orthomcl.org/orthomcl/; MAFFT (v 7.2.71), https://mafft.cbrc.jp/alignment/software/; TreeBeST (v 1.9.2), http://treesoft.source-forge.net/treebest.shtml; HyPhy, https://veg.github.io/hyphy-site/; Datamonkey, http://www.datamonkey.org/; STAR, http://star.mit.edu/genetics/; featureCounts, http://bioinf.wehi.edu.au/feature-Counts/; DESeq2, https://bioconductor.org/packages/release/bioc/html/DESeq2.html; SARTools, https://github.com/PF2-pasteur-fr/SARTools/; Dotter, https://sonnhammer.sbc.su.se/Dotter.html; GATK (v 3.3-0-g37228af), https://software.broadinstitute.org/gatk/; KAT comp, https://github.com/TGAC/KAT/; BUSCO (v 2), http://busco.ezlab.org/; Trimmomatic (v 0.36 PE), http://www.usadellab.org/cms/?page=trimmomatic; Bowtie2 (v 2.2.4), http://bowtie-bio.sourceforge.net/bowtie2/index.shtml; MACS2 (v 2.0.10.20131216), https://github.com/taoliu/MACS/; R (v 3.2.5), https://www.r-project.org/; gplots (v 3.0.1), https://cran.r-project.org/web/packages/gplots/index.html; bedtools (v 2.25.0), http://bedtools.readthedocs.io/en/latest/; kSamples (v 1.2-4), https://cran.r-project.org/web/packages/kSamples/index.html; ggbiplot (v 0.55), https://github.com/vqv/ggbiplot/; Tandem Repeats Finder, https://tandem.bu.edu/trf/trf.html; seqLogo, https://bioconductor.org/packages/release/bioc/html/seqLogo.html; RNAfold, http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi; UniProt/Swiss-Prot, http://www.uniprot.org/; dammit!, https://dammit.readthedocs.io/en/refactor-1.0/; Transfuse, https://github.com/cboursnell/transfuse/; GMAP, http://research-pub.gene.com/gmap/; Trim Galore!, https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/; Kallisto, https://pachterlab.github.io/kallisto/; Sleuth, https://pachterlab.github.io/sleuth_walkthroughs/trapnell/analysis.html; All-vsl-all BLASTP (version 2.2.30+), https://blast.ncbi.nlm.nih.gov/Blast.cgi; MUSCLE (v 3.8.31), https://www.drive5.com/muscle/; HMMER suit (v 3.1b1 May 2013), http://hmmer.org/; FASTASEARCH (v 36.8.8), https://www.ebi.ac.uk/Tools/sss/fasta/; Integrative Genomics Viewer (IGV) (v 2.3.97), https://github.com/ssadedin/IGV-CRAM/; MEGA (v 7.0.18), https://www.megasoftware.net/; RAxML (v 8.2.11), https://sco.h-its.org/exelixis/web/software/raxml/index.html; Burrows-Wheeler aligner (v 0.7.15), http://bio-bwa.sourceforge.net/; Samtools (v 1.3), http://www.htslib.org/; Geneious (v 10.2.3), https://www.geneious.com/; Coancestry, https://www.zsl.org/science/software/coancestry/; PLINK (v 1.07), http://zzz.bwh.harvard.edu/plink/.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0153-5.

References

1. Meredith, R. W., Krajewski, C., Westerman, M. & Springer, M. S. Relationships and divergence times among the orders and families of Marsupialia. *Mus. North. Ariz. Bull.* **65**, 383–406 (2009).
2. Black, K. H., Price, G. J., Archer, M. & Hand, S. J. Bearing up well? Understanding the past, present and future of Australia's koalas. *Gondwana Res.* **25**, 1186–1201 (2014).
3. Gleadow, R. M., Haburjak, J., Dunn, J. E., Conn, M. E. & Conn, E. E. Frequency and distribution of cyanogenic glycosides in *Eucalyptus* L'Hérit. *Phytochemistry* **69**, 1870–1874 (2008).
4. Nagy, K. & Martin, R. Field metabolic rate, water flux, food consumption and time budget of koalas, *Phascolarctos cinereus* (Marsupialia: Phascolarctidae) in Victoria. *Aust. J. Zool.* **33**, 655–665 (1985).
5. Woinarski, J. C., Burbidge, A. A. & Harrison, P. L. Ongoing unraveling of a continental fauna: decline and extinction of Australian mammals since European settlement. *Proc. Natl. Acad. Sci. USA* **112**, 4531–4540 (2015).
6. Adams-Hosking, C. et al. Use of expert knowledge to elicit population trends for the koala (*Phascolarctos cinereus*). *Divers. Distrib.* **22**, 249–262 (2016).
7. McAlpine, C. et al. Conserving koalas: a review of the contrasting regional trends, outlooks and policy challenges. *Biol. Conserv.* **192**, 226–236 (2015).
8. Martin, R. & Handasyde, K. A. *The Koala: Natural History, Conservation and Management.* (UNSW Press: Sydney, New South Wales, Australia (1999).
9. Hrdina, F. & Gordon, G. The koala and possum trade in Queensland, 1906–1936. *Aust. Zool.* **32**, 543 (2004).
10. Menkhorst, P. Hunted, marooned, re-introduced, contracepted: a history of koala management in Victoria. in *Too Close for Comfort: Contentious Issues in Human–Wildlife Encounters* (eds. Lunney, D. et al.) 73–92 (Royal Zoological Society of NSW, Mosman, New South Wales, Australia, 2008).
11. Seymour, A. M. et al. High effective inbreeding coefficients correlate with morphological abnormalities in populations of South Australian koalas (*Phascolarctos cinereus*). *Anim. Conserv.* **4**, 211–219 (2001).
12. Simmons, G., Clarke, D., McKee, J., Young, P. & Meers, J. Discovery of a novel retrovirus sequence in an Australian native rodent (*Melomys burtoni*): a putative link between gibbon ape leukemia virus and koala retrovirus. *PLoS One* **9**, e106954 (2014).
13. Alfano, N. et al. Endogenous gibbon ape leukemia virus identified in a rodent (*Melomys burtoni* subsp.) from Wallacea (Indonesia). *J. Virol.* **90**, 8169–8180 (2016).
14. Tarlinton, R. E., Meers, J. & Young, P. R. Retroviral invasion of the koala genome. *Nature* **442**, 79–81 (2006).
15. Xu, W. et al. An exogenous retrovirus isolated from koalas with malignant neoplasias in a US zoo. *Proc. Natl. Acad. Sci. USA* **110**, 11547–11552 (2013).
16. Taylor-Brown, A. & Polkinghorne, A. New and emerging chlamydial infections of creatures great and small. *New Microbes New Infect.* **18**, 28–33 (2017).
17. Hayman, D. Marsupial cytogenetics. *Aust. J. Zool.* **37**, 331–349 (1989).
18. Deakin, J. E. et al. Anchoring genome sequence to chromosomes of the central bearded dragon (*Pogona vitticeps*) enables reconstruction of ancestral squamate macrochromosomes and identifies sequence content of the Z chromosome. *BMC Genomics* **17**, 447 (2016).
19. Brown, J.D. & O'Neill, R.J. The evolution of centromeric DNA sequences. *Encyclopedia of Life Sciences* https://doi.org/10.1002/9780470015902.a0020827.pub2 (Wiley, Hoboken, NJ, USA, 2014).
20. Carone, D. M. et al. A new class of retroviral and satellite encoded small RNAs emanates from mammalian centromeres. *Chromosoma* **118**, 113–125 (2009).
21. Earnshaw, W. C. & Rothfield, N. Identification of a family of human centromere proteins using autoimmune sera from patients with scleroderma. *Chromosoma* **91**, 313–321 (1985).
22. O'Neill, R. J. W., O'Neill, M. J. & Graves, J. A. M. Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. *Nature* **393**, 68–72 (1998).
23. Nagaki, K. et al. Sequencing of a rice centromere uncovers active genes. *Nat. Genet.* **36**, 138–145 (2004).
24. Zhang, Y. et al. Structural features of the rice chromosome 4 centromere. *Nucleic Acids Res.* **32**, 2023–2030 (2004).
25. Carbone, L. et al. Centromere remodeling in *Hoolock leuconedys* (Hylobatidae) by a new transposable element unique to the gibbons. *Genome Biol. Evol.* **4**, 648–658 (2012).
26. Grant, J. et al. Rsx is a metatherian RNA with Xist-like properties in X-chromosome inactivation. *Nature* **487**, 254–258 (2012).
27. Hobbs, M. et al. A transcriptome resource for the koala (*Phascolarctos cinereus*): insights into koala retrovirus transcription and sequence diversity. *BMC Genomics* **15**, 786 (2014).
28. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
29. Foley, W. J. & Moore, B. D. Plant secondary metabolites and vertebrate herbivores–from physiological regulation to ecosystem function. *Curr. Opin. Plant Biol.* **8**, 430–435 (2005).
30. Eschler, B. M., Pass, D. M., Willis, R. & Foley, W. J. Distribution of foliar formylated phloroglucinol derivatives amongst Eucalyptus species. *Biochem. Syst. Ecol.* **28**, 813–824 (2000).
31. Pass, G. J., McLean, S., Stupans, I. & Davies, N. Microsomal metabolism of the terpene 1,8-cineole in the common brushtail possum (*Trichosurus vulpecula*), koala (*Phascolarctos cinereus*), rat and human. *Xenobiotica* **31**, 205–221 (2001).
32. Ngo, S. N. T., McKinnon, R. A. & Stupans, I. Cloning and expression of koala (*Phascolarctos cinereus*) liver cytochrome P450 CYP4A15. *Gene* **376**, 123–132 (2006).
33. Myburg, A. A. et al. The genome of *Eucalyptus grandis*. *Nature* **510**, 356–362 (2014).
34. Kirischian, N., McArthur, A. G., Jesuthasan, C., Krattenmacher, B. & Wilson, J. Y. Phylogenetic and functional analysis of the vertebrate cytochrome P450 2 family. *J. Mol. Evol.* **72**, 56–71 (2011).
35. Nelson, D. R. The cytochrome P450 homepage. *Hum. Genomics* **4**, 59–65 (2009).
36. Miners, J. O. & Birkett, D. J. Cytochrome P4502C9: an enzyme of major importance in human drug metabolism. *Br. J. Clin. Pharmacol.* **45**, 525–538 (1998).
37. Davies, N. M. & Skjodt, N. M. Clinical pharmacokinetics of meloxicam. A cyclo-oxygenase-2 preferential nonsteroidal anti-inflammatory drug. *Clin. Pharmacokinet.* **36**, 115–126 (1999).
38. Kimble, B. et al. In vitro hepatic microsomal metabolism of meloxicam in koalas (*Phascolarctos cinereus*), brushtail possums (*Trichosurus vulpecula*), ringtail possums (*Pseudocheirus peregrinus*), rats (*Rattus norvegicus*) and dogs (*Canis lupus familiaris*). *Comp. Biochem. Physiol. C Toxicol. Pharmacol.* **161**, 7–14 (2014).
39. Blanshard, W. & Bodley, K. Koalas. in *Medicine of Australian Mammals* (eds. Vogelnest, L. & Woods, R.) 307–327 (Csiro Publishing, Melbourne, Victoria, Australia, 2008).
40. Villalba, J. J., Provenza, F. D. & Bryant, J. Consequences of the interaction between nutrients and plant secondary metabolites on herbivore selectivity: benefits or detriments for plants? *Oikos* **97**, 282–292 (2002).
41. Kratzing, J. E. The anatomy and histology of the nasal cavity of the koala (*Phascolarctos cinereus*). *J. Anat.* **138**, 55–65 (1984).
42. Moore, B. D., Foley, W. J., Wallis, I. R., Cowling, A. & Handasyde, K. A. Eucalyptus foliar chemistry explains selective feeding by koalas. *Biol. Lett.* **1**, 64–67 (2005).
43. Freeland, W.J. & Janzen, D.H. Strategies in herbivory by mammals: the role of plant secondary compounds. *Am. Nat.* **108**, 269–289 https://doi.org/10.1086/282907 (1974).
44. McBride, C. S. Rapid evolution of smell and taste receptor genes during host specialization in *Drosophila sechellia*. *Proc. Natl. Acad. Sci. USA* **104**, 4996–5001 (2007).
45. Watson, K. J. et al. Expression of aquaporin water channels in rat taste buds. *Chem. Senses* **32**, 411–421 (2007).
46. Rosen, A. M., Roussin, A. T. & Di Lorenzo, P. M. Water as an independent taste modality. *Front. Neurosci.* **4**, 175 (2010).
47. Gilbertson, T. A., Baquero, A. F. & Spray-Watson, K. J. Water taste: the importance of osmotic sensing in the oral cavity. *J. Water Health* **4**, 35–40 (2010).
48. Meyerhof, W. et al. The molecular receptive ranges of human TAS2R bitter taste receptors. *Chem. Senses* **35**, 157–170 (2010).
49. Hayakawa, T., Suzuki-Hashido, N., Matsui, A. & Go, Y. Frequent expansions of the bitter taste receptor gene repertoire during evolution of mammals in the Euarchontoglires clade. *Mol. Biol. Evol.* **31**, 2018–2031 (2014).
50. Li, D. & Zhang, J. Diet shapes the evolution of the vertebrate bitter taste receptor gene repertoire. *Mol. Biol. Evol.* **31**, 303–309 (2014).
51. Li, R. et al. The sequence and de novo assembly of the giant panda genome. *Nature* **463**, 311–317 (2010).
52. Johnston, S. D., McGowan, M. R., O'Callaghan, P., Cox, R. & Nicolson, V. Studies of the oestrous cycle, oestrus and pregnancy in the koala (*Phascolarctos cinereus*). *J. Reprod. Fertil.* **120**, 49–57 (2000).
53. Morris, K. M. et al. Characterisation of the immune compounds in koala milk using a combined transcriptomic and proteomic approach. *Sci. Rep.* **6**, 35011 (2016).
54. Department of the Environment. *Phascolarctos cinereus* (combined populations of Queensland, New South Wales and the Australian Capital Territory) in Species Profile and Threats Database (Department of the Environment, Canberra, Australian Capital Territory, 2016).
55. Polkinghorne, A., Hanger, J. & Timms, P. Recent advances in understanding the biology, epidemiology and control of chlamydial infections in koalas. *Vet. Microbiol.* **165**, 214–223 (2013).

56. Rhodes, J. R. et al. Using integrated population modelling to quantify the implications of multiple threatening processes for a rapidly declining population. *Biol. Conserv.* **144**, 1081–1088 (2011).

57. Morris, K. et al. The koala immunological toolkit: sequence identification and comparison of key markers of the koala (*Phascolarctos cinereus*) immune response. *Aust. J. Zool.* **62**, 195–199 (2014).

58. Morris, K. M. et al. Identification, characterisation and expression analysis of natural killer receptor genes in *Chlamydia pecorum* infected koalas (*Phascolarctos cinereus*). *BMC Genomics* **16**, 796 (2015).

59. Cheng, Y. et al. Characterisation of MHC class I genes in the koala. *Immunogenetics* **70**, 125–133 (2018).

60. Jones, E. A., Cheng, Y., O'Meally, D. & Belov, K. Characterization of the antimicrobial peptide family defensins in the Tasmanian devil (*Sarcophilus harrisii*), koala (*Phascolarctos cinereus*), and tammar wallaby (*Macropus eugenii*). *Immunogenetics* **69**, 133–143 (2017).

61. Burton, M. J. et al. Pathogenesis of progressive scarring trachoma in Ethiopia and Tanzania and its implications for disease control: two cohort studies. *PLoS Negl. Trop. Dis.* **9**, e0003763 (2015).

62. Derrick, T., Roberts, C., Last, A. R., Burr, S. E. & Holland, M. J. Trachoma and ocular chlamydial infection in the era of genomics. *Mediators Inflamm.* **2015**, 791847 (2015).

63. Stoye, J. P. Koala retrovirus: a genome invasion in real time. *Genome Biol.* **7**, 241 (2006).

64. Hobbs, M. et al. Long-read genome sequence assembly provides insight into ongoing retroviral invasion of the koala germline. *Sci. Rep.* **7**, 15838 (2017).

65. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).

66. Roberts, R. G. et al. New ages for the last Australian megafauna: continent-wide extinction about 46,000 years ago. *Science* **292**, 1888–1892 (2001).

67. Field, J., Wroe, S., Trueman, C. N., Garvey, J. & Wyatt-Spratt, S. Looking for the archaeological signature in Australian megafaunal extinctions. *Quat. Int.* **285**, 76–88 (2013).

68. Neaves, L. E. et al. Phylogeography of the koala, (*Phascolarctos cinereus*), and harmonising data to inform conservation. *PLoS One* **11**, e0162207 (2016).

69. Tsangaras, K. et al. Historically low mitochondrial DNA diversity in koalas (*Phascolarctos cinereus*). *BMC Genet.* **13**, 92 (2012).

70. Taylor, A. C., Graves, J. A., Murray, N. D. & Sherwin, W. B. Conservation genetics of the koala (*Phascolarctos cinereus*). II. Limited variability in minisatellite DNA sequences. *Biochem. Genet.* **29**, 355–363 (1991).

71. Taylor, A. C. et al. Conservation genetics of the koala (*Phascolarctos cinereus*): low mitochondrial DNA variation amongst southern Australian populations. *Genet. Res.* **69**, 25–33 (1997).

72. Dennison, S. et al. Population genetics of the koala (*Phascolarctos cinereus*) in north-eastern New South Wales and south-eastern Queensland. *Aust. J. Zool.* **64**, 402–412 (2017).

73. Cristescu, R. et al. Inbreeding and testicular abnormalities in a bottlenecked population of koalas (*Phascolarctos cinereus*). *Wildl. Res.* **36**, 299–308 (2009).

74. Frankham, R. et al. Predicting the probability of outbreeding depression. *Conserv. Biol.* **25**, 465–475 (2011).

75. Frankham, R. et al. *Genetic Management of Fragmented Animal and Plant Populations* (Oxford University Press, Oxford, 2017).

76. Hansen, J., Sato, M., Russell, G. & Kharecha, P. Climate sensitivity, sea level and atmospheric carbon dioxide. *Philos. Trans. A Math. Phys. Eng. Sci.* **371**, 20120294 (2013).

77. O'Connell, J. F. & Allen, J. The process, biotic impact, and global implications of the human colonization of Sahul about 47,000 years ago. *J. Archaeol. Sci.* **56**, 73–84 (2015).

78. Clarkson, C. et al. Human occupation of northern Australia by 65,000 years ago. *Nature* **547**, 306–310 (2017).

79. Saltré, F. et al. Climate change not to blame for late Quaternary megafauna extinctions in Australia. *Nat. Commun.* **7**, 10511 (2016).

80. Wang, J. Triadic IBD coefficients and applications to estimating pairwise relatedness. *Genet. Res.* **89**, 135–153 (2007).

81. Wang, J. COANCESTRY: a program for simulating, estimating and analysing relatedness and inbreeding coefficients. *Mol. Ecol. Resour.* **11**, 141–145 (2011).

82. Warren, W. C. et al. Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453**, 175–183 (2008).

83. Mikkelsen, T. S. et al. Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature* **447**, 167–177 (2007).

84. Renfree, M. B. et al. Genome sequence of an Australian kangaroo, *Macropus eugenii*, provides insight into the evolution of mammalian reproduction and development. *Genome Biol.* **12**, R81 (2011).

85. Murchison, E. P. et al. Genome sequencing and analysis of the Tasmanian devil and its transmissible cancer. *Cell* **148**, 780–791 (2012).

Acknowledgements

Author contributions

R.N.J., K.B., P.T. and M.R.W. designed the original concept and scientific objectives and oversaw the project and analyses. R.N.J., D.C., M.D.B.E., A.K.G., D.O., A.K. and P.T. acquired samples for sequencing. T.L.R., M.R.W., Z.C., D.O., G.J.E. and F.D.P. performed library preparation, genome sequencing, assembly and annotation. S.Y.W.H. performed PSMC analysis. A.Y.-Y.C. characterized repetitive sequences. R.J.O., T.N.H. and Z.D. characterized centromeric and telomeric regions. C.M.W. and M.B.R. annotated and analyzed reproductive and developmental genes. K.M.M. annotated and analyzed lactation genes. T.H. and D.C. annotated and analyzed *TAS1R* and *TAS2R* genes. H.R.P. annotated and analyzed *OR* genes. D.C. annotated and analyzed aquaporin genes. K.B., Y.C., P.A.B., E.A.J., D.O. and E.P. annotated and analyzed *MHC*, *Ig*, *TCR*, *NK* and antimicrobial genes. A.P., K.B., D.O. and D.M. analyzed the ocular RNA-seq data. P.A.B., B.W., C.E.G., P.T., K.B. and A.P investigated candidate genes for chlamydia vaccine response. P.T., M.R.W., R.S., M.H., A.K., A.D.G. and K.T. characterized retrovirus insertions and wrote the KoRV sections of the manuscript. J.G.B., S.D., M.D.B.E., G.J.F., L.E.N., R.N.J., B.W. and C.J.H. contributed to analyses and interpretation of exon capture sequence data. P.D.W., S.A.W. and H.R.P. annotated and analyzed *RSX* data. W.J.N., C.E.G., Y.C., W.H., F.D., M.G., K.M.E., B.W. and C.R. analyzed *CYP* genes. C.E.G. and C.M.W. analyzed *V1R* genes. J.E.D., A.G. and H.R.P. constructed super-scaffolds. J.A.M.G., V.A., F.D., C.J.H., K.M.H., A.P., B.W., D.C., M.H., D.E.A.-P., P.A.B., L.E.N., C.E.G., S.A.W. and C.E.H. provided constructive feedback on data analysis and interpretation. R.N.J., P.T., K.B., M.R.W., A.P., M.D.B.E. and G.J.F. obtained funding and other resources. R.N.J. and K.B. wrote the manuscript with input from all other authors.

Competing interests

The authors declare no competing interests.

Additional information

Methods

General methods. A full description of the Methods can be found in the Supplementary Note. No statistical methods were used to predetermine sample size.

Genome sequencing and assembly of the koala reference genome. *Sequencing.* Samples were obtained as part of veterinary care at the Port Macquarie Koala Hospital and Australia Zoo Wildlife Hospital, and from the Australian Museum Tissue Collection. Sample collection was performed in accordance with methods approved by the Australian Museum Animal Ethics Committee (permit numbers 11–03 and 15–05). "Pacific Chocolate" (Australian Museum registration M.45022), a female from Port Macquarie in northeast New South Wales, was sampled immediately after euthanasia by veterinary staff at the Port Macquarie Koala Hospital (27 June 2012), following unsuccessful treatment of severe chlamydiosis. Two koalas from southeast Queensland—a female, "Bilbo" (Australian Museum registration M.47724), from Upper Brookfield, and a male, "Birke", from Birkdale— were sampled following euthanasia due to severe chlamydiosis (20 August 2015) and severe injuries (26 August 2012), respectively. High molecular weight (HMW) DNA was extracted from heart tissue for Pacific Chocolate and kidney tissue for Birke using the DNeasy Blood and Tissue kit (Qiagen), with RNaseA (Qiagen) treatment. HMW DNA from Bilbo was extracted for PacBio sequencing from spleen tissue using Genomic-Tip 100/G columns (Qiagen), DNA Buffer set (Qiagen) and RNaseA (Qiagen) treatment. Fifteen SMRTbell libraries were prepared (RCG) as per the PacBio 20-kb template preparation protocol, with an additional damage repair step performed after size selection. A minimum size cutoff of 15 or 20 kb was used in the size selection stage using the Sage Science BluePippin system. The libraries were sequenced on the Pacific Biosciences RS II platform (Pacific Biosciences) employing P6 C4 chemistry with either 240 min or 360 min movie lengths. A total of 272 SMRT Cells were sequenced to give an estimated overall coverage of 57.3× based on a genome size of 3.5 Gbp. A TruSeq DNA PCR free library was constructed with a mean library insert size of 450 bp. 400,473,997 paired-end reads were generated yielding a minimum coverage of 34×. HMW gDNA was sequenced on an Illumina 150bpPE HiSeq X Ten sequencing run (Illumina)

Assembly. An overlapping layout consensus assembly algorithm, FALCON (v 0.3.0) (see URLs), was used to generate the draft genome using PacBio reads. Total genome coverage before assembly was estimated by total bases from reads divided by 3.5 Gbp genome size. The estimated total coverage is 57.3×. FALCON leverages error-corrected long seed reads to generate an overlapping layout consensus representation of the genome. Approximately 23× of long reads are required by FALCON as seed reads, and the rest are used for error correction. The seed read length of the reads at the 60% percentile was calculated as 10,889 bp. The FALCON assembly was run on Amazon Web Service Tokyo region using r3.8xlarge spot instances as compute node, with the number of instances varying from 12 to 20 depending on availability.

After filtering low-quality and duplicate reads, approximately 57.3-fold long-read coverage was used for assembly. The primary contigs from the FALCON v 0.3.0 assembly (representing homozygous regions of the genome) yielded genome version phaCin_unsw_v4.1. This comprised 3.19 Gb, including 1,906 contigs with an N50 of 11.6 Mb and sizes ranging up to 40.6 Mb. The heterozygous regions of the genome (representing the alternative contigs from the assembly) were a total of 230 Mb, with an N50 of 48.8 Kb (Supplementary Table 2). Approximately 30-fold coverage of Illumina short reads was used to polish the assembly with Pilon[86].

BUSCO analysis on the draft assembly was run against the mammalian ortholog database with the –long parameter on all genomes under comparison. This initial analysis showed the assembly only reached about 60% of genome completeness, suggesting a high number of indels in the draft genome. The genome polishing tool Pilon[86] was employed to improve draft assembly from FALCON. About 30× of 150 bp paired-end Illumina X Ten short reads from Bilbo was used as an input for this polishing process, which was run on a compute cluster provided by Intersect Australia Limited.

We implemented the method of Deakin et al.[18] for super-scaffolding. Briefly, tables of homologous genes were generated using the physical order of genes on the chromosomes of gray short-tailed opossum and tammar wallaby as references and koala phaCin_unsw_v4.1 (Bilbo) as target (Supplementary Table 4).

Analysis of centromeric regions and repeat structure. Repeat content was called using RepeatMasker with combined RepBase libraries (v 2015-08-07) and RepeatModeller calls generated from the genome assemblies. The resulting calls were then filtered using custom Python scripts to remove short fragments (see "Code availability") and combine tandem or overlapping repeat calls. To characterize the centromeric regions of the genome, chromatin immunoprecipitation (ChIP) was performed using the Invitrogen MAGnify Chromatin Immunoprecipitation System (Revision 6). Repeat content of the centromeric regions was determined using RepBase annotated marsupial repeats and output from RepeatModeller analysis of koala. RepeatMasker was used to locate repeats. Candidate centromeric segments were identified using two sliding window analyses, with window sizes of 200 kb and 20 kb and step sizes

of 100 kb and 10 kb, respectively. Small tandem repeats were discovered in koala *RSX* sequence using the Tandem Repeat Finder program[87], using +2, −3, and −7 as scores for match, mismatch and gap opening, respectively. Alignments of consensus repeat units with the *RSX* sequence were processed to obtain nucleotide frequency at each position.

Genome annotation and gene family analysis. Annotations were generated using the automated genome annotation pipeline MAKER[88,89]. We masked repeats in the assembly by providing MAKER with a koala-specific repeat library generated with RepeatModeler[90], against which RepeatMasker (v 4.0.3)[91] queried genomic contigs. Gene annotations were made using a protein database combining the UniProt/ Swiss-Prot[92] protein database, all sequences for human (*Homo sapiens*), gray short-tailed opossum (*Monodelphis domestica*), Tasmanian devil (*Sarcophilus harrisii*) and tammar wallaby (*Notamacropus eugenii*) from the NCBI protein database[93], and a curated set of marsupial and monotreme immune genes[94]. We downloaded all published koala mRNAseq reads from SRA (PRJNA230900, PRJNA327021) and reassembled de novo male, female and mammary transcriptomes using the default parameters of Trinity v 2.3.2[95]. Each assembly was filtered such that contigs accounting for 90% of mapped reads were passed to MAKER as homologous transcript evidence. Ab initio gene predictions were made using the programs SNAP[96], Genemark[97] and Augustus[98]. Three iterative runs of MAKER were used to produce the final gene set.

Gene families were called using NCBI Blast (2.3.0) OrthoMCL (2.0.9)[99]. The protein sequences of genes belonging to orthogroups identified by OrthoMCL were aligned using MAFFT (7.2.71)[100] and the gene tree was inferred using TreeBeST (1.9.2)[101] providing a species tree to guide the phylogenetic reconstruction. Custom scripts (see "Code availability") were applied to identify families with expansion within the koala, Diprotodontia, Australidelphia and marsupial lineages.

Sequence evolution. Sequence evolution on specific gene families was conducted on the cytochrome P450 (*CYP*), vomeronasal receptor (*V1R*), olfactory receptor (*OR*), aquaporin and taste receptor genes (Supplementary Note). Genes involved in koala development and reproduction and lactation were also characterized (Supplementary Note). Koala *MHC*, *TCR* and *IGG* genes were annotated and analyzed for expression between diseased and healthy animals (Supplementary Note). Evidence of selection across *CYP* and *V1R* genes was evaluated (Supplementary Note) using multispecies alignments (*N* = 152 and 8 sequences, respectively) in HyPhy[102], hosted by the Datamonkey webserver[103].

RNA-seq analysis of koala conjunctival tissue samples. Conjunctival tissue samples were collected from 26 koalas euthanized due to injury or disease by veterinarians at Australia Zoo Wildlife Hospital, Currumbin Wildlife Hospital and Moggill Koala Hospital. The collection protocol was approved by the University of the Sunshine Coast Animal Ethics Committee (AN/S/15/36). Health assessments of the eye were performed by an experienced veterinarian and classified as either 'healthy' (*N* = 13) or 'diseased' (*N* = 13) based on evidence of gross pathology consistent with ocular chlamydiosis[55]. Conjunctival tissue samples from each animal were placed directly in RNALater (Qiagen, Germany) buffer overnight at 4 °C before storing at −80 °C for later use. RNA was extracted using an RNeasy Mini Kit (Qiagen, Germany) according to the manufacturer's instructions, with an on-column DNase treatment to eliminate contaminating DNA from the sample. The concentration and quality of the isolated RNA was determined using a NanoDrop ND-1000 160 Spectrophotometer and Agilent BioAnalyzer (Agilent, USA). Library construction and sequencing were performed by the Ramaciotti Centre (UNSW, Kensington, NSW) with TruSeq stranded mRNA chemistry on a NextSeq500 (Illumina, USA). Reads were mapped to the phCin_unsw_v4.1 assembly using the default parameters of STAR[104] and counts summed over features using featureCounts[105]. Differentially expressed genes were called using DESeq2[106] as implemented in the SARTools package[107].

Koala retrovirus (KoRV). We searched for KoRV sequences within the scaffolds of the phaCin_unsw v4.1 assembly of the Bilbo genome sequence, and also within alternative contig sequences before their correction by Pilon (since we noticed that in a few cases KoRV sequences were removed in the course of the sequence polishing process). KoRV sequences were found by using the program blastn[108] to search with KoRV genome reference sequences (GenBank AF151794 and AB721500). Search results were converted to BED format and the KoRV and recKoRV components of each read were merged with the program mergeBed. KoRV insertions within genes were identified using the program intersectBed[109]. Pre-integration allelic sequences were found by using blastn[108] to search the phaCin_unsw v4.1 genome sequence assembly with sequences flanking KoRV/ recKoRV integrations as queries. In two cases the expected allelic sequence was not present in the Bilbo genome, but was found by searching the genome of another koala (Pacific Chocolate). To check the expected relationship between pairs of allelic sequences, we inspected dot plot alignments of representative sequences (not shown) created with the program dotter[110].

Koala population genomics: historical population size. Demographic history was inferred from the diploid sequence of each of the three koalas, using a

pairwise sequential Markovian coalescent (PSMC) method[65]. We conducted a range of preliminary analyses and found that PSMC plots were not sensitive to the values chosen for the maximum number of iterations (N), the number of free atomic time intervals (p), the maximum time to the most recent common ancestor (t), and the initial value of ρ. Based on these investigations, our final PSMC analyses of the three genome sequences used values of $N = 25$, $t = 5$, $\rho = 1$ and $p = 4 + 25 \times 2 + 4 + 6$. The number of atomic time intervals is similar to that recommended for analyses of modern human genomes[65], which are similar in size to the koala genomes. We determined the variance in estimates of N_e using 100 bootstrap replicates. Replicate analyses in which we varied the values of p, t and ρ produced PSMC plots that were broadly similar to those using our chosen 'optimal' settings (Supplementary Fig. 10).

The plots of demographic history were scaled using a generation length of 7 years, corresponding to the midpoint of the range of 6 to 8 years estimated for the koala[111] and the midpoints of the estimates of the human mutation rate (1.45×10^{-8} mutations per site per generation; summarized by ref. [112]) and mouse mutation rate (5.4×10^{-9} mutations per site per generation[113]) were applied in the absence of a mutation rate estimate for koala (Supplementary Fig. 10). The koala mutation rate is likely to be closer to that of humans, based on greater similarity in genome size, life history, and effective population size, relative to mouse[112].

Koala population genomics: contemporary population analysis. Forty-nine koalas were sampled throughout the distribution using a hierarchical approach to allow examination of genetic relationships at a range of scales, from familial to range-wide. All individuals were sequenced using a target capture approach described in ref. [114], with a kit targeting 2,167 marsupial exon sequences. Illumina sequence reads were quality-filtered and trimmed (see ref. [114] for details) and mapped to the koala genome (Bowtie2, v2.2.4[115]). A panel of 4,257 SNP sites was identified (using GATK version 3.3-0-g37228af[116]) that showed expected levels of relatedness and differentiation among the sampled individuals. A panel of 1,200 SNPs (obtained by mapping to targets, filtering, and selecting one SNP per target) showed fine-scale regional differentiation consistent with evolutionary history and recent population management (Fig. 3).

Statistics and reproducibility. In Fig. 1e, points shown indicate the mean empirical Bayes factor (EBF) for sites under selection; error bars, 95% confidence interval. In Fig. 1f–h, 95% confidence intervals are calculated as $1.96 \times$ s.e.m. (sample size is sequence depth, as indicated by red bars in Fig. 1c).

In Fig. 3c, center lines indicate median and box limits indicate upper and lower quartiles. Upper whisker = min(max(x), Q_3 + 1.5 × IQR), lower whisker = max(min(x), Q_1 − 1.5 × IQR); i.e., upper whisker = upper quartile + 1.5 × box length, lower whisker = lower quartile − 1.5 × box length. Circles indicate outliers. Linear modeling indicated that mean F differed significantly between several regions (Midcoast New South Wales–Southern Australia, $P = 0.000524$; Queensland–Southern New South Wales, $P = 0.00237$; Queensland–Southern Australia, $P = 0.00000107$; Southeast Queensland–Southern Australia, $P = 0.006596$).

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Code availability. (1) Custom scripts to identify gene families with expansion within the koala, Diprotodontia, Australidelphia and marsupial lineages; (2) custom scripts to identify refined repeat calls; and (3) code used to generate SNP genotypes from exon capture data are available at https://github.com/DrRebeccaJ/KoalaGenome.

Data availability. The *Phascolarctos cinereus* BioSamples are as follows: Bilbo 61053, SAMN06198159; Pacific Chocolate, SAMEA91939168; Birke. SAMEA103910665. Koala Genome Consortium Projects for the Koala Whole Genome Shotgun project and genome assembly are registered under the umbrella BioProject PRJEB19389 (union of PRJEB5196 and PRJNA359763).

Transcriptome data are submitted under PRJNA230900 (adrenal, brain, heart, lung, kidney, uterus, liver and spleen) and PRJNA327021 (milk and mammary gland). Illumina short-read data for Birke is submitted under PRJEB19982.

The Bilbo 61053 assembly described in this paper is version MSTS01000000 and consists of sequences MSTS01000001–MSTS01001906. For the Bilbo assembly Illumina X Ten reads are submitted under PRJEB19457 and PacBio reads under PRJEB19889.

ChIP-seq data have been deposited under BioProject PRJNA415832 and GEO GSE111153.

References

86. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
87. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
88. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
89. Yandell, M. & Ence, D. A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* **13**, 329–342 (2012).
90. Smit, A., Hubley, R. & Green, P. RepeatModeler Open-1.0. 2008–2015 (2014).
91. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0. 2013–2015 (2015).
92. Boutet, E. et al. UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. in *Plant Bioinformatics: Methods and Protocols* (ed. Edwards, D.) 23–54 (2016).
93. O'Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
94. Wong, E. S., Papenfuss, A. T. & Belov, K. Immunome database for marsupials and monotremes. *BMC Immunol.* **12**, 48 (2011).
95. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
96. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
97. Borodovsky, M. & Lomsadze, A. Gene identification in prokaryotic genomes, phages, metagenomes, and EST sequences with GeneMarkS suite. *Curr. Protoc. Bioinformatics* **4**, 4.5.1–4.5.17 (2011).
98. Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
99. Li, L., Stoeckert, C. J. Jr & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
100. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
101. Vilella, A. J. et al. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* **19**, 327–335 (2009).
102. Pond, S.L.K. & Muse, S.V. HyPhy: hypothesis testing using phylogenies. in *Statistical Methods in Molecular Evolution* 125–181 (Springer, New York, 2005).
103. Delport, W., Poon, A. F., Frost, S. D. & Kosakovsky Pond, S. L. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* **26**, 2455–2457 (2010).
104. Dobin, A. & Gingeras, T. R. Mapping RNA-seq reads with STAR. *Curr. Protoc. Bioinformatics* **11**, 11.14.1–11.14.19 (2015).
105. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
106. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
107. Varet, H., Brillet-Guéguen, L., Coppée, J.-Y. & Dillies, M.-A. SARTools: a DESeq2- and edgeR-based R pipeline for comprehensive differential analysis of RNA-Seq data. *PLoS One* **11**, e0157022 (2016).
108. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
109. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
110. Sonnhammer, E. L. & Durbin, R. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167**, GC1–GC10 (1995).
111. Phillips, S. S. Population trends and the koala conservation debate. *Conserv. Biol.* **14**, 650–659 (2000).
112. Lynch, M. et al. Genetic drift, selection and the evolution of the mutation rate. *Nat. Rev. Genet.* **17**, 704–714 (2016).
113. Uchimura, A. et al. Germline mutation rates and the long-term phenotypic effects of mutation accumulation in wild-type laboratory mice and mutator mice. *Genome Res.* **25**, 1125–1134 (2015).
114. Bragg, J. G., Potter, S., Bi, K. & Moritz, C. Exon capture phylogenomics: efficacy across scales of divergence. *Mol. Ecol. Resour.* **16**, 1059–1068 (2016).
115. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
116. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

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 <u>exact sample size</u> (*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 <u>central tendency</u> (e.g. median, mean) and <u>variation</u> (e.g. standard deviation, interquartile range)
☒	☐	Clearly defined error bars in <u>all</u> 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.

naturereasearch

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.