Published in final edited form as:
Nat Genet. 2018 August ; 50(8): 1102–1111. doi:10.1038/s41588-018-0153-5.

Adaptation and conservation insights from the koala genome

A full list of authors and affiliations appears at the end of the article.

Abstract

The koala is the only extant species of the marsupial family Phascolarctidae and is now classified as 'vulnerable' due to habitat loss and widespread disease. We sequenced the koala genome, producing the most complete and contiguous marsupial reference genome to date. We show that the koala's ability to detoxify eucalypt foliage, toxic to most other mammals, 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 characterised centromeres and novel lactation proteins that protect young in the pouch, as well as immune responses to chlamydial disease. Historical demography revealed a significant population crash coincident with the decline of Australian megafauna, while contemporary populations revealed biogeographic boundaries and increased inbreeding in populations impacted by historic translocations. Genetically diverse populations requiring habitat corridors and translocation programs were identified and provide the key to the koala's survival in the wild.

*corresponding author: Rebecca.Johnson@austmus.gov.au.
^contributed equally
#these authors jointly supervised this work

Author Contributions:

R.N.J., K.B., P.T., M.R.W. designed the original concept, scientific objectives and oversaw the project and analyses.
R.N.J., D.C., M.D.B.E., A.G.K., D.O., A.K., P.T. acquired samples for sequencing.
T.L.R., M.R.W., Z.C., D.O., G.J.E., 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 analysed reproductive and developmental genes.
K.M.M. annotated and analysed lactation genes.
T.H. and D.C. annotated and analysed *TAS1R* and *TAS2R* genes.
H.R.P. annotated and analysed *OR* genes.
D.C. annotated and analysed *Aquaporin* genes.
K.B., Y.C., P.A.B., E.A.J., D.O., E.P. annotated and analysed *MHC, Ig, TCR, NK* and antimicrobial genes.
A.P., K.B., D.O., D.M. analysed the ocular RNASeq data.
P.A.B., B.W., C.E.G., P.T., K.B., A.P investigated candidate genes for chlamydia vaccine response.
P.T., M.R.W., R.S., M.H., A.K., A.D.G., K.T. characterised 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., C.J.H. contributed to analyses and interpretation of exon capture sequence data.
P.D.W., S.A.W., H.R.P. annotated and analysed *RSX* data.
W.J.N., C.E.G., Y.C., W.H., F.D., M.G., K.M.E., B.W., C.R. analysed *CYP* genes.
C.E.G., C.M.W. analysed *V1R* genes.
J.E.D., A.G., 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.A.P., P.A.B., L.E.N., C.E.G., S.A.W., 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., G.J.F. obtained funding and other resources.
R.N.J., K.B. wrote the manuscript with input from all other authors.

Competing financial interests statement:

The authors declare no competing financial interests.

Europe PMC Funders Author Manuscripts

Introduction

The koala is an iconic Australian marsupial, instantly recognisable by its round, humanoid face and distinctive body shape. Fossil evidence reveals 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 underdeveloped young. Birth occurs after just 35 days 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 mother's 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 5. Due to the low calorific content of this diet, the koala rests and sleeps up to 22 hours a day 6. A detailed understanding of the mechanisms by which koalas detoxify eucalyptus and protect their young in the pouch has previously eluded us, 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 euthanize 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 7, and koala numbers have plummeted in parts of the northern end of its range since European settlement of the continent 8, but increased in sections of the southern end 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 9. The species was heavily exploited by a pelt trade (1870s to late 1920s) which harvested millions of animals 8,10,11. Today, the threats are primarily due to loss and fragmentation of habitat, urbanisation, 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 8. Koalas present a complex conservation conundrum: causes of decline in the north include due to ongoing habitat fragmentation, urbanisation, and disease; yet in the south has followed a different path 12, with widespread, often sequential, translocations (1920-1990s) from a limited founder population have resulted in genetically bottlenecked populations that are overabundant, to the point of starvation, in some areas 13. 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 arrived in Australia, it is postulated, via a putative murine vector before cross-species transmission 14,15. It is now widespread in northern koalas and appears to be spreading to southern populations 16. Some strains appear to be more virulent than others and are putatively associated with an increase in neoplastic disease 17. 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 18. A complete koala genome offers insights into the species' genetic susceptibility to these diseases, provides the genomic basis for innovative vaccines, and can underpin novel conservation management solutions incorporating 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 19. 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 1906 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) totalled 230 Mb, with an N50 of 48.8 kb (Table 1; Methods and Supplementary Tables 1-3). 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 24 to assemble long-read contigs into 'virtual' chromosome scaffolds (or 'super-contigs') (Supplementary Tables 4-5 and Supplementary Note 2.1). 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 "black holes" in eutherian genome assemblies due to intractable higher order arrays of satellites (e.g. human and mouse) 27. Centromeres are also smaller in marsupials than eutherians, so more amenable to analysis 28. ChIP-seq using centromeric antibodies (CENP-A and CREST) 29 enabled the identification of scaffolds containing putative centromeric regions (Supplementary Fig. 3) and characterisation of known and novel repeats, including composite elements within koala centromeric domains (Supplementary Table 6; also Supplementary Tables 7-10) yet lack the previously annotated retroelement, Kangaroo Endogenous Retrovirus (KERV), found in some tammar wallaby centromeres 30. 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 other species with small centromeres 27,28,31,32, koala centromeres lack higher-order satellite arrays (Supplementary Tables 7-10). Among the novel repeats we identified some are similar to composite elements recently described in gibbon centromeres 33, in which absence of higher order satellite arrays accompanied the evolution of novel 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 28,32,33.

Interspersed repeats account for approximately 47.5% of the koala genome, 44% of which are transposable elements (Supplementary Table 11). As in other mammalian genomes,

SINEs and 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 characterisation and annotation of repeat-rich long non-coding RNAs, including *Rsx* which mediates X chromosome inactivation in female marsupials 34. Koala *Rsx* represents the first marsupial *Rsx* to be fully annotated and to have its structure predicted (Supplementary Fig. 4 and Supplementary Note 2.4). As expected, it was expressed in all female tissues, but in no male tissues 37.

The assembled koala genome has very high coverage of coding regions: we recovered 95.1% of 4,104 mammalian BUSCOs 38, the highest for any published marsupial genome (Supplementary Table 5) and comparable with 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 characterised the genome, we undertook detailed analyses of key genes and gene families in order to gain insights into the genomic basis of the koala's highly specialised biology. Gene families of particular interest were those that encode proteins involved in induced ovulation, in the complex lactation process, those responsible for immunity, and those enzymes that enable the koala to subsist on a toxic diet.

Unique biology of the koala

Ability to tolerate a highly toxic diet

The koala's diet of eucalyptus leaves contains high levels of plant secondary metabolites (PSMs) 39 phenolic compounds 40 and terpenes (e.g. 41) that would be lethal to most other mammals 42. Unsurprisingly, koalas experience little competition for food resources. *Eucalyptus grandis*, showed substantial expansion in terpene synthase genes relative to other plant genomes 43. Eucalypt toxicity is therefore likely to have exerted intense selection pressure on the koala's ability to metabolise 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 haem-thiolate enzymes that play a role in detoxification through phase 1 oxidative metabolism of a range of compounds including xenobiotics 44. These genes have been identified throughout the tree of life: including in plants, animals, fungi, bacteria and viruses 45. 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, revealing particularly high expression in the liver, consistent with a role in detoxification (Supplementary Fig. 6).

Comparing *CYP2C* gene context in mouse versus koala revealed 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 (Fig. 1c; Supplementary Note 3.3), while purifying selection shapes the rest of the gene (Fig. 1c; 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 neofunctionalisation within the *CYP*s. Such adaptations enable koalas to detoxify their highly specialised and PSM-rich diet.

The characterisation of koala *CYP2Cs* has significant therapeutic potential. The high expression levels of *CYP2C* genes in the liver explain why meloxicam, a non-steroidal anti-inflammatory drug (NSAID) known to be metabolised by *CYP2C* in humans 46,47, and frequently used for pain relief in veterinary care, is so rapidly metabolised in the koala and a handful of other eucalypt-eating marsupials (common brush-tail possum and eastern ring-tail possum) compared with eutherian species 47,48. It is expected that other NSAIDs are also rapidly metabolised in koalas and have little efficacy at currently suggested doses 49. Anti-*Chlamydia* antibiotics like chloramphenicol are degraded rapidly by koalas; treatment with a single dose applicable for humans is insufficient in koalas, which require a daily dose for up to 30 to 45 days. 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 PSMs 50. Koalas have been observed to sniff leaves before tasting them 51, and their acute discrimination has been correlated with the complexity and concentration of PSMs 52. This suggests an important role for olfaction and vomerolfaction, as well as taste. While most herbivores circumvent plant chemical defences by detoxifying one or a few compounds 53, the complexity of eucalyptus PSMs, in combination with the terpene expansion in eucalypts, led us to hypothesise that the koala requires superior capabilities both in specialist detection and in PSM detoxification. We therefore investigated the genomic basis of the koala's taste and smell senses, and found multiple gene family expansions that could enhance its ability to make food choices.

Here we report an expansion of one lineage of vomeronasal receptor type 1 (*V1R*) genes associated with the detection of non-volatile odorants (Supplementary Note 3.4). There are six in koala, compared with 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 between diverse PSMs.

Surprisingly, given the degree of its dietary specialisation, the olfactory receptor (OR) genes characterised in koala (1,169 OR genes) revealed a gene repertoire 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 3.5). This may be understood in the context of relaxed selection on olfactory receptors among dietary specialists 54.

We also report genomic evidence of expansions within the taste receptor families that would enable the koala to optimise ingestion of leaves with a higher moisture and nutrient content in concert with the concentration of toxic PSMs in their food plants. The koala's ability to 'taste water' is likely to be enhanced by an apparent functional duplication of the aquaporin 5 gene 55–57 (Supplementary Table 14; Supplementary Note 3.6).

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 PSMs 5,40,41,58. In marsupials the *TAS2R* family includes the orthologous repertoires from eutherians, and in addition three specific expansions in the last common ancestor shared by all marsupials 59,60 (Fig. 2). Massive 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 amongst the most of all mammal species 59,60, including paralogs of human and mouse receptors whose agonists are toxic glycosides (Supplementary Note 3.7; Supplementary Table 15). The *TAS1R* gene families, responsible for sweet taste and umami amino acid perception, have previously been reported as pseudogenized in eutherians with highly specialised diets, such as the giant panda 61. In the koala, however, we found that all *TAS1R* genes are putatively functional (Supplementary Fig. 7).

Genomics of an induced ovulator

Koala reproduction is particularly interesting because the koala is an induced ovulator 62. The genome contains all of the 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 3.8). We also 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 3.8), possibly to prevent sperm leakage from the female reproductive tract in this arboreal species.

Genomic characterisation of koala milk

A koala young is about the size of a kidney bean and weighs <0.5 gram. It crawls into the mother's backward-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 characterise the main components of koala milk (Supplementary Fig. 8;) (Supplementary Table 16; Supplementary Note 3.9; and 63). The high-quality assembly of the genome enabled both identification of marsupial-specific genes, and determination of their evolutionary origins based on their genomic locations. For instance, we found that four *LLP* genes are tightly linked to both Trichosurin and Beta-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.

Meanwhile, koala Marsupial Milk 1 (*MM1*) 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 54 (Supplementary Fig. 8). This region in eutherians 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. We showed that Phci-CATH5 has broad-spectrum antimicrobial activity against a range of bacteria and fungi (unpublished data E.P., Y.C., D.O., K.B.) and is able to significantly reduce the infectivity of *Chlamydia pecorum* by rapidly inactivating elementary bodies prior to infection, and is thus a potential topical agent for the treatment of ocular chlamydiosis (unpublished data E.P., Y.C., D.O., K.B.).

Koala immunome and disease

At the time of European settlement koalas were widespread in eastern mainland Australia, from north Queensland to the south-eastern 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* 64. There is strong evidence to suggest that some fragmented populations of koalas are already facing extinction, particularly in formerly densely populated koala territories in south-east Queensland and northern NSW. A major challenge for the conservation of these declining koala populations is the high prevalence of disease, especially caused by the obligate intracellular bacterial pathogen, *Chlamydia pecorum*, which is found across the range, with the exception of some offshore islands 65. A primary challenge for managing these populations has been the lack of knowledge about the koala immune response to disease. Recent modelling suggests the best way to stabilise heavily affected koala populations is to target disease 66.

The long-read-based genome enabled the *de novo* assembly of complex, highly duplicated immune gene families and the most comprehensive annotation of immune gene clusters in any marsupial 63,67,68. These include the Major Histocompatibility Complex (*MHC*) 69, as well as T-Cell Receptors (*TCR*), immunoglobulin (*IG*) (Supplementary Fig. 9; Supplementary Table 17-18; and Supplementary Note 3.10), Natural Killer cell (NK) receptor 68 and defensin 70 gene clusters. Together these findings provide a springboard for new disease research and allow us to interrogate the immune response to the most significant pathogen of the koala: *Chlamydia pecorum.*

Of the more than 1000 koalas presenting annually to wildlife hospitals in Queensland and NSW, 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). Basic case/control association tests for five koalas involved in a chlamydia vaccination trial revealed the MHCII *DMA* and *DMB* genes, as well as the *CD8-a* gene, may be involved in differential immune responses to chlamydia vaccine (Supplementary Note 3.11, Supplementary Table 21). We also conducted differential expression analysis of RNASeq data from conjunctival tissue collected from koalas at necropsy, both with and

without signs of ocular chlamydiosis, revealing that in diseased animals 1508 of the 26558 annotated genes (5.7%) were two-fold upregulated, while 685 (2.6%) were downregulated by greater than two-fold when compared with healthy animals (Supplementary Note 3.12; and Supplementary Fig. 9). In diseased animals, upregulated genes were associated with 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 Th1 rather than Th2 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 71. Furthermore, resolution of human trachoma infection is thought to require a IFN-γ driven Th1 response 72, and in diseased koalas we found that IFN-γ was upregulated 4.7-fold in the conjunctival tissue. These annotated koala immune genes – the first data on the mucosal immune response to chlamydial disease – will now enable 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) 73, which is spreading from the north of the country to the south. Both endogenous (germline transmission) and exogenous (infectious “horizontal” transmission) forms are extant 74. Our results provide the first comprehensive view of KoRV insertions in a koala genome. We found a total of 73 insertions in the phaCin_unsw_4.1 assembly (detailed in 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 23-24).

We investigated the sites of KoRV insertion to define their proximity to protein-coding genes and explore possible disruptions. This revealed 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 LTR could possibly affect the transcription of the host genes.

Understanding the genetics of host resistance to chlamydia and the aetiology 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 generalised across bio-regions, and further complicated by unique regional issues described above. Since it is not possible to generalise management, it is imperative that decisions are informed by empirical data relevant to each bio-region.

The koala genome allowed the unique opportunity to combine historical evolutionary data with high-resolution contemporary population genomic markers in order to address these management challenges. To infer the ancient demographic history of the species, we used the long-read reference genome and short read data from two other koalas, using the

pairwise sequentially Markovian coalescent (PSMC) method 75 (Fig. 3a) (Supplementary Fig. 10; and Methods). The data show that the modern koala, which appeared in the fossil record 350kya 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 84 and 30,000-40,000 years ago 85. The koala demonstrates that there was ongoing survival of at least some species present at the time 85.

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 mtDNA80,86 although over a shorter time scale.

We analysed populations of recent koala samples using 1200 SNPs derived from targeted capture libraries mapped to the koala genome (Supplementary Note 5.2). We found significant levels of genetic diversity with limited fine-scale differentiation consistent with long-term connectivity across regions. We find clear evidence of low genetic diversity in southern koalas, consistent with a recent history of sequential translocations 10,80,87,88 (Fig. 3b). At a continental scale, we reveal biogeographic barriers to gene flow associated with the Brisbane Valley and Clarence River as identified by mtDNA studies 80,81 and reveal a previously undetected 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 NSW and Queensland currently show high levels of genetic diversity.

The information generated here provides a critical foundation for conservation management to maintain gene flow regionally whilst incorporating the genetic legacy of biogeographic barriers. Furthermore, the stark 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 89.

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 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 whilst minimizing risks 90,91. Future utilization 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 considerable insights into historic and contemporary population dynamics, providing essential 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, it 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. Some of this work is already underway. As members of government advisory committees, some of the authors have initiated inclusion of genomic information into the NSW 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 colonisation 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 off-shore islands, so risk assessment should be carried out before embarking on translocations. Sequencing the genome has significantly 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.

Koala Genome Online Methods

A full description of the Methods can be found in the Supplementary Information. 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, 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 (27/06/2012), 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 (20/08/2015) and severe injuries (26/8/2012) 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) added following digestion. HMW DNA from "Bilbo" was extracted from spleen tissue of using Genomic-Tip 100/G

columns (Qiagen) and DNA Buffer set (Qiagen). Fifteen SMRTbell libraries were prepared (RCG) as per the PacBio 20kb template preparation protocol, with an additional damage repair step performed after size selection. A minimum size cutoff of 15 or 20kb was utilized 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 minor 360 min movie lengths. A total of 272 SMRT Cells were sequenced to give an estimated overall coverage of 57.3x based on a genome size of 3.5Gbp. 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 34x. 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

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.sourceforge.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/featureCounts/)
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)

before assembly was estimated by total bases from reads divided by 3.5 Gbp genome size. The estimated total coverage is 57.3x. FALCON leverages error-corrected long seed reads to generate an overlapping layout consensus representation of the genome. Approximately 23x 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 12~20 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 1906 contigs with an N50 of 11.6 Mb and sizes ranging up to 40.6 Mb. The heterozygous regions of the genome (representing the alternate contigs from the assembly) were a total of 230 Mb, with an N50 of 48.8 Kb (Supplementary Table 2.1). Approximately 30-fold coverage of Illumina short reads was used to polish the assembly with Pilon 92.

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 92, was employed to improve draft assembly from FALCON. About 30x 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 24 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, 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

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/)

centromeric segments were identified using two sliding window analyses, with a window size of 200 kb and 20 kb and a step size of 100 kb and 10 kb respectively. Small tandem repeats were discovered in koala *RSX* sequence using the Tandem Repeat Finder program 93, 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. Chip-seq data is deposited under Bioproject: PRJNA415832 and GEO submission: GSE111153 (see URLs).

Genome annotation and gene family analysis

Annotations were generated using the automated genome annotation pipeline MAKER 94,95. We masked repeats in the assembly by providing MAKER with a koala specific repeat library generated with RepeatModeler 96, against which *RepeatMasker* (v 4.0.3) 97 queried genomic contigs. Gene annotations were made using a protein database combining the Uniprot/Swiss-Prot 98, protein database and 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 99, and a curated set of marsupial and monotreme immune genes 100. 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 2.3.2 101. 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 102, Genemark 103, and Augustus 104. 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, 105). The protein sequences of genes belonging to orthogroups identified by OrthoMCL were aligned using MAFFT (7.2.71, 106) and the gene tree was inferred using TreeBest (1.9.2, 107) providing a species tree to guide the phylogenetic reconstruction. Custom scripts 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*) (Supplementary Note 3.2-3.3), vomeronasal receptor (*V1R*) genes (Supplementary Note 3.4), Olfactory Receptor (*OR*) genes (Supplementary Note 3.5), Aquaporins (Supplementary Note 3.6),Taste receptor genes (Supplementary Note 3.7). Genes involved in koala development and reproduction (Supplementary Note 3.8), and lactation (Supplementary Note 3.9) were also characterised. Koala MHC, TCR and IGG genes were annotated and analysled for expression between diseaed and healthy animals (Supplementary Note 3.10-3.11). Evidence of selection across *CYP* and *V1R* genes was evaluated (Supplementary Note 3.3-3.4) using multispecies alignments (N = 152 and 8 sequences, respectively) in HyPhy 108, hosted by datamonkey webserver 109.

RNASeq analysis of koala conjunctival tissue samples

Conjunctival tissue samples were collected from 26 koalas euthanised 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 65. Conjunctival tissue samples from each animal were placed directly in RNALater (Qiagen, Germany) buffer overnight at 4°C prior to 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 BioAnalyser (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 110 and counts summed over features using featureCounts 111. Differentially expressed genes were called using DESeq2 112 as implemented in the SARTools package 113. Reads have been deposited in the SRA under the accession BioProject PRJEB19389.

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 prior to 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 using by using the program blastn 114 to search with KoRV genome reference sequences (GenBank AF151794 and AB721500) as well as with a recKoRV sequence from Löber et al. 115. 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 116. Pre-integration allelic sequences were found by using blastn 114 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 dotplot alignments of representative sequences (not shown) created with the program dotter 117.

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 75. 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 rho (r). Based on these investigations, our final PSMC

analyses of the three genome sequences used values of N=25, t=5, r=1, and p=4+25*2+4+6. The number of atomic time intervals is similar to that recommended for analyses of modern human genomes 75, 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 r 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 118 and the midpoint of the estimates of the human mutation rate (1.45×10^{-8} mutations per site per generation; summarised by 119) and mouse mutation rate (5.4×10^{-9} mutations per site per generation 120) was 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 119.

Contemporary population analysis—Fifty-six 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 121, with a kit targeting 2167 marsupial exon sequences. Illumina sequence reads were quality-filtered and trimmed (see 74 for details) and mapped to the koala genome (Bowtie2, v2.2.4 122). A panel of 4257 SNP sites was identified (using GATK version 3.3-0-g37228af 123) that showed expected levels of relatedness and differentiation among the sampled individuals.

A set of SNP sites was identified (using GATK version 3.3-0-g37228af 123) and showed expected levels of relatedness among individual samples. A panel of 1200 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).

Code Availability Statement

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

Data Availability Statement

The *Phascolarctos cinereus* BioSamples are as follows: Bilbo 61053 - SAMN06198159, Pacific Chocolate - SAMEA91939168 and 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 is submitted under PRJNA230900 (adrenal, brain, heart, lung, kidney, uterus, liver and spleen) and PRJNA327021 (milk and mammary gland). Chip-seq data have

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

Statistics and Reproducibility

Figure 1e: Points shown in panel e indicate the mean Empirical Bayes Factor EBF for sites underselection; error bars 95% confidence interval. In panels f, g, and h, 95% confidence intervals are calculated as 1.96 × SEM (sample size is sequence depth, as indicated by red bars in panel a).

Figure 3c: Centre 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); ie upper whisker = upper quartile + 1.5 * box length, lower whisker = lower quartile − 1.5 * box length. Circles indicate outliers. Linear modelling 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, $P = 0.006596$).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

Rebecca N. Johnson[*,^,#,1,2], Denis O'Meally[^,2,3], Zhiliang Chen[^,4], Graham J. Etherington[5], Simon Y. W. Ho[2], Will J. Nash[5], Catherine E. Grueber[2,6], Yuanyuan Cheng[2,8], Camilla M. Whittington[2], 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[5,6], 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[9], Alex D. Greenwood[17,18], Takashi Hayakawa[19,20], Kristofer M. Helgen[1,21], Matthew Hobbs[1], Clare E. Holleley[22], Thomas N. Heider[10], Elizabeth A. Jones[9], Andrew King[1], Danielle Madden[3], Jennifer A. Marshall Graves[14,23,24], Katrina M. Morris[25], Linda E. Neaves[1,26], Hardip R. Patel[12], 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[^,4,11], Peter Timms[^,29], and Katherine Belov[^,#,2]

Europe PMC Funders Author Manuscripts

Affiliations

[1]Australian Museum Research Institute, Australian Museum, Sydney NSW, Australia

[2]School of Life and Environmental Sciences, Faculty of Science, University of Sydney, NSW, Australia

[3]Animal Research Centre, Faculty of Science, Health, Education & Engineering, University of the Sunshine Coast, Maroochydore, QLD, Australia

[4]School of Biotechnology and Biomolecular Sciences, University of New South Wales, Kensington, NSW, Australia

[5]Earlham Institute, Norwich Research Park, Norwich, United Kingdom

[6]Wellcome Trust Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom

[7]San Diego Zoo Global, San Diego, CA, USA

[8]UQ Genomics Initiative, University of Queensland, QLD, Australia

[9]Sydney School of Veterinary Science, Faculty of Science, University of Sydney, Sydney, NSW, Australia

[10]Department of Molecular and Cell Biology and Institute for Systems Genomics, University of Connecticut, Storrs, CT, USA

[11]Ramaciotti Centre for Genomics, University of New South Wales, Kensington, NSW, Australia

[12]John Curtin School of Medical Research, Australian National University, Acton, ACT, Australia

[13]National Herbarium of NSW, Royal Botanic Gardens & Domain Trust, Sydney, NSW, Australia

[14]Institute for Applied Ecology, University of Canberra, Bruce, ACT, Australia

[15]Department of Biological Sciences, University of East Anglia, Norwich, United Kingdom

[16]Australia Zoo Wildlife Hospital, Beerwah, QLD, 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, Aichi, Japan

[20]Japan Monkey Centre, Inuyama, Aichi 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, SA, Australia

[22]Australian National Wildlife Collection, National Research Collections Australia, CSIRO. Canberra ACT, Australia

[23]Research School of Biology, Australian National University, Canberra, ACT, Australia

[24]School of Life Sciences, La Trobe University, Bundoora, Vic, Australia

[25]The Roslin Institute and R(D)SVS, University of Edinburgh, Easter Bush, Midlothian, United Kingdom

[26]Royal Botanic Garden Edinburgh, Edinburgh, United Kingdom

[27]School of BioSciences, University of Melbourne, Melbourne, VIC, 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, QLD, Australia

Acknowledgements

R.N.J. and the Australian Museum acknowledge the Australian Museum Foundation, BioPlatforms Australia, New South Wales Environmental Trust grant 2014/RD/0015, the University of Sydney HPC service and Amazon Web Services for support; and Chad Staples from Featherdale Wildlife Park, Cheyne Flanagan from Port Macquarie Koala Hospital, Jon Hangar, Emily Hynes, Jackie Reed, Sandy Ingleby, Anja Divljan, and Scott Ginn, for assistance with sample acquisition. K.B. acknowledges support from the Australian Research Council and Bioplatforms Australia. M.R.W. and the Ramaciotti Centre for Genomics acknowledge support from the Australian Research Council, from the Australian Government NCRIS scheme via Bioplatforms Australia, the New South Wales State Government RAAP scheme and the University of New South Wales. W.H. and W.J.H. were supported by strategic BBSRC funding (Institute Strategic Programme Grant BB/J004669/1) and by the NBI Computing Infrastructure for Science (CiS) group. A.D.G., K.M.H and K.T. were supported by Grant Number R01GM092706 from the National Institute of General Medical Sciences (NIGMS) and A.D.G. had additional support from the Morris Animal Foundation Grant Number D14ZO-94. T.N.H., Z.R.D. and R.J.O. were supported by awards from the National Science Foundation 1613806 and the facilities within the Center for Genome Innovation at the University of Connecticut. C.H. thanks CSIRO National Research Collections Australia funding. K.B. and A.P. thank the veterinary staff at Australia Zoo Wildlife Hospital, Currumbin Wildlife Hospital and Moggill Koala Hospital for their assistance in the collection of samples for the koala conjunctival transcriptome study. T.H. acknowledges the Kyoto University Research Administration Office (KURA) for support and was financed by the JSPS KAKENHI Grant Number 16K18630 and the Sasakawa Scientific Research Grant from the Japan Science Society. A.P. and P.T. acknowledge financial support from the Australian Research Council and A.G. for financial support via Australian Research Council Discovery Grant DP110104377. CMW is supported by a University of Sydney research fellowship from the estate of Mabs Melville. All authors thank BioPlatforms Australia and Pacific Biosciences. The authors thank three anonymous referees and Timothy Haydon for valuable editorial input on the manuscript; Sally Potter for expert technical assistance; and Ros Gleadow, Celine Frere, Dan Lunney and David Alvarez-Ponce for valuable discussions on content.

References

1. Meredith RW, Krajewski C, Westerman M, Springer MS. Relationships and divergence times among the orders and families of Marsupialia. Museum of Northern Arizona Bulletin. 2009; 65:383–406.
2. Black KH, Price GJ, Archer M, Hand SJ. Bearing up well? Understanding the past, present and future of Australia's koalas. Gondwana Research. 2014; 25:1186–1201.
3. Munemasa M, et al. Phylogenetic analysis of diprotodontian marsupials based on complete mitochondrial genomes. Genes & Genetic Systems. 2006; 81:181–191. [PubMed: 16905872]
4. May-Collado LJ, Kilpatrick CW, Agnarsson I. Mammals from ‘down under’: a multi-gene species-level phylogeny of marsupial mammals (Mammalia, Metatheria). PeerJ. 2015; 3:e805. [PubMed: 25755933]

5. Gleadow RM, Haburjak J, Dunn J, Conn M, Conn EE. Frequency and distribution of cyanogenic glycosides in Eucalyptus L'Hérit. Phytochemistry. 2008; 69:1870–1874. [PubMed: 18474385]
6. Nagy K, Martin R. Field Metabolic Rate, Water Flux, Food Consumption and Time Budget of Koalas, *Phascolarctos Cinereus* (Marsupialia: Phascolarctidae) in Victoria. Australian Journal of Zoology. 1985; 33:655–665.
7. Woinarski JC, Burbidge AA, Harrison PL. Ongoing unraveling of a continental fauna: decline and extinction of Australian mammals since European settlement. Proceedings of the National Academy of Sciences. 2015; 112:4531–4540.
8. Adams-Hosking C, et al. Use of expert knowledge to elicit population trends for the koala (*Phascolarctos cinereus.*). Diversity and Distributions. 2016; 22:249–262.
9. McAlpine C, et al. Conserving koalas: a review of the contrasting regional trends, outlooks and policy challenges. Biological Conservation. 2015; 192:226–236.
10. Martin R, Handasyde KA. The koala: natural history, conservation and management. UNSW Press; 1999.
11. Hrdina F, Gordon G. The koala and possum trade in Queensland, 1906-1936. Australian Zoologist. 2004; 32:543.
12. Menkhorst P. Hunted, marooned, re-introduced, contracepted: a history of Koala management in VictoriaToo Close for Comfort: Contentious Issues in Human–Wildlife Encounters. Lunney D, Munn A, Meikle W, editorsRoyal Zoological Society of NSW; Mosman, NSW: 2008. 73–92.
13. Seymour AM, et al. High effective inbreeding coefficients correlate with morphological abnormalities in populations of South Australian koalas (*Phascolarctos cinereus*). Animal Conservation. 2001; 4:211–219.
14. 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. 2014; 9:e106954. [PubMed: 25251014]
15. Alfano N, et al. Endogenous gibbon ape leukemia virus identified in a rodent (*Melomys burtoni* subsp.) from Wallacea (Indonesia). Journal of Virology. 2016; 90:8169–8180. [PubMed: 27384662]
16. Tarlinton RE, Meers J, Young PR. Retroviral invasion of the koala genome. Nature. 2006; 442:79–81. [PubMed: 16823453]
17. Xu W, et al. An exogenous retrovirus isolated from koalas with malignant neoplasias in a US zoo. Proceedings of the National Academy of Sciences. 2013; 110:11547–11552.
18. Taylor-Brown A, Polkinghorne A. New and emerging chlamydial infections of creatures great and small. New Microbes and New Infections. 2017; 18:28. [PubMed: 28560043]
19. Hayman D. Marsupial cytogenetics. Australian Journal of Zoology. 1989; 37:331–349.
20. Warren WC, et al. Genome analysis of the platypus reveals unique signatures of evolution. Nature. 2008; 453:175–183. [PubMed: 18464734]
21. Mikkelsen TS, et al. Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. Nature. 2007; 447:167–177. [PubMed: 17495919]
22. Renfree MB, et al. Genome sequence of an Australian kangaroo, *Macropus eugenii*, provides insight into the evolution of mammalian reproduction and development. Genome Biology. 2011; 12:1.
23. Murchison EP, et al. Genome sequencing and analysis of the Tasmanian devil and its transmissible cancer. Cell. 2012; 148:780–791. [PubMed: 22341448]
24. Deakin JE, 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. 2016; 17
25. Rens W, et al. Reversal and convergence in marsupial chromosome evolution. Cytogenetic and Genome Research. 2004; 102:282–290.
26. Deakin J, Graves J, Rens W. The evolution of marsupial and monotreme chromosomes. Cytogenetic and Genome Research. 2012; 137:113–129. [PubMed: 22777195]
27. Brown JD, O'Neill RJ. The Evolution of Centromeric DNA Sequences. eLS. 2014

28. Carone DM, et al. A new class of retroviral and satellite encoded small RNAs emanates from mammalian centromeres. Chromosoma. 2009; 118:113–125. [PubMed: 18839199]
29. Earnshaw WC, Rothfield N. Identification of a family of human centromere proteins using autoimmune sera from patients with scleroderma. Chromosoma. 1985; 91:313–321. [PubMed: 2579778]
30. O'Neill RJW, O'Neill MJ, Graves JAM. Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. Nature. 1998; 393:68. [PubMed: 9590690]
31. Nagaki K, et al. Sequencing of a rice centromere uncovers active genes. Nature Genetics. 2004; 36:138–145. [PubMed: 14716315]
32. Zhang Y, et al. Structural features of the rice chromosome 4 centromere. Nucleic Acids Research. 2004; 32:2023–2030. [PubMed: 15064362]
33. Carbone L, et al. Centromere remodeling in *Hoolock leuconedys* (Hylobatidae) by a new transposable element unique to the gibbons. Genome Biology and Evolution. 2012; 4:760–770.
34. Grant J, et al. Rsx is a metatherian RNA with Xist-like properties in X-chromosome inactivation. Nature. 2012; 487:254–258. [PubMed: 22722828]
35. Deakin JE, et al. Reconstruction of the ancestral marsupial karyotype from comparative gene maps. BMC Evolutionary Biology. 2013; 13:258. [PubMed: 24261750]
36. Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL. The vienna RNA websuite. Nucleic Acids Research. 2008; 36:W70–W74. [PubMed: 18424795]
37. Hobbs M, et al. A transcriptome resource for the koala (*Phascolarctos cinereus*): insights into koala retrovirus transcription and sequence diversity. BMC Genomics. 2014; 15:1. [PubMed: 24382143]
38. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015 btv351.
39. Foley WJ, Moore BD. Plant secondary metabolites and vertebrate herbivores – from physiological regulation to ecosystem function. Current Opinion in Plant Biology. 2005; 8:430–435. [PubMed: 15939665]
40. Eschler B, Pass D, Willis R, Foley W. Distribution of foliar formylated phloroglucinol derivatives amongst Eucalyptus species. Biochemical Systematics and Ecology. 2000; 28:813–824. [PubMed: 10913843]
41. Pass G, 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. 2001; 31:205–221. [PubMed: 11465406]
42. Ngo SNT, McKinnon RA, Stupans I. Cloning and expression of koala (*Phascolarctos cinereus*) liver cytochrome P450 CYP4A15. Gene. 2006; 376:123–132. [PubMed: 16677781]
43. Myburg AA, et al. The genome of *Eucalyptus grandis*. Nature. 2014; 510:356–362. [PubMed: 24919147]
44. Kirischian N, McArthur AG, Jesuthasan C, Krattenmacher B, Wilson JY. Phylogenetic and functional analysis of the vertebrate cytochrome P450 2 family. Journal of Molecular Evolution. 2011; 72:56–71. [PubMed: 21116621]
45. Nelson DR. The cytochrome p450 homepage. Human Genomics. 2009; 4:59. [PubMed: 19951895]
46. Miners JO, Birkett DJ. Cytochrome P4502C9: an enzyme of major importance in human drug metabolism. British Journal of Clinical Pharmacology. 1998; 45:525–538. [PubMed: 9663807]
47. Davies NM, Skjodt NM. Clinical pharmacokinetics of meloxicam. Clinical Pharmacokinetics. 1999; 36:115–126. [PubMed: 10092958]
48. 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*). Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology. 2014; 161:7–14. [PubMed: 24345479]
49. Blanshard W, Bodley K. KoalasMedicine of Australian Mammals. Vogelnest L, Woods R, editorsCSIRO PUBLISHING; Melbourne: 2008. 307–327.

50. Villalba JJ, Provenza FD, Bryant J. Consequences of the interaction between nutrients and plant secondary metabolites on herbivore selectivity: benefits or detriments for plants? Oikos. 2002; 97:282–292.
51. Kratzing JE. The anatomy and histology of the nasal cavity of the koala (*Phascolarctos cinereus*). Journal of Anatomy. 1984; 138:55. [PubMed: 6706839]
52. Moore BD, Foley WJ, Wallis IR, Cowling A, Handasyde KA. Eucalyptus foliar chemistry explains selective feeding by koalas. Biology Letters. 2005; 1:64–67. [PubMed: 17148129]
53. Freeland WJ, Janzen DH. Strategies in herbivory by mammals: the role of plant secondary compounds. American Naturalist. 1974:269–289.
54. McBride CS. Rapid evolution of smell and taste receptor genes during host specialization in Drosophila sechellia. Proceedings of the National Academy of Sciences. 2007; 104:4996–5001.
55. Watson KJ, et al. Expression of aquaporin water channels in rat taste buds. Chemical Senses. 2007; 32:411–421. [PubMed: 17339611]
56. Rosen AM, Roussin AT, Di Lorenzo PM. Water as an independent taste modality. Frontiers in Neuroscience. 2010; 4
57. Gilbertson TA, Baquero AF, Spray-Watson KJ. Water taste: the importance of osmotic sensing in the oral cavity. Journal of Water and Health. 2006; 4:35–40. [PubMed: 16493898]
58. Meyerhof W, et al. The molecular receptive ranges of human TAS2R bitter taste receptors. Chemical Senses. 2010; 35:157–170. [PubMed: 20022913]
59. 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. Molecular Biology and Evolution. 2014; 31:2018–2031. [PubMed: 24758778]
60. Li D, Zhang J. Diet shapes the evolution of the vertebrate bitter taste receptor gene repertoire. Molecular Biology and Evolution. 2014; 31:303–309. [PubMed: 24202612]
61. Li R, et al. The sequence and de novo assembly of the giant panda genome. Nature. 2010; 463:311–317. [PubMed: 20010809]
62. Johnston S, McGowan M, O'Callaghan P, Cox R, Nicolson V. Studies of the oestrous cycle, oestrus and pregnancy in the koala (*Phascolarctos cinereus*). Journal of Reproduction and Fertility. 2000; 120:49–57. [PubMed: 11006145]
63. Morris KM, et al. Characterisation of the immune compounds in koala milk using a combined transcriptomic and proteomic approach. Scientific Reports. 2016; 6 35011.
64. Department of the Environment. Species Profile and Threats Database. *Phascolarctos cinereus* (combined populations of Queensland, New South Wales and the Australian Capital Territory)Department of the Environment; Canberra, Australian Capital Territory: 2016.
65. Polkinghorne A, Hanger J, Timms P. Recent advances in understanding the biology, epidemiology and control of chlamydial infections in koalas. Veterinary Microbiology. 2013; 165:214–223. [PubMed: 23523170]
66. Rhodes JR, et al. Using integrated population modelling to quantify the implications of multiple threatening processes for a rapidly declining population. Biological Conservation. 2011; 144:1081–1088.
67. Morris K, et al. The koala immunological toolkit: sequence identification and comparison of key markers of the koala (*Phascolarctos cinereus*) immune response. Australian Journal of Zoology. 2014; 62:195–199.
68. Morris KM, et al. Identification, characterisation and expression analysis of natural killer receptor genes in Chlamydia pecorum infected koalas (*Phascolarctos cinereus*). BMC Genomics. 2015; 16:796. [PubMed: 26471184]
69. Cheng Y, et al. Characterisation of MHC class I genes in the koala. Immunogenetics. 2017:1–9. [PubMed: 27933432]
70. Jones EA, 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. 2017; 69:133–143. [PubMed: 27838759]
71. Burton MJ, et al. Pathogenesis of progressive scarring trachoma in Ethiopia and Tanzania and its implications for disease control: two cohort studies. PLoS Neglected Tropical Diseases. 2015; 9:e0003763. [PubMed: 25970613]

Europe PMC Funders Author Manuscripts

72. Derrick T, Last AR, Burr SE, Holland MJ. Trachoma and ocular chlamydial infection in the era of genomics. Mediators of Inflammation. 2015; 2015
73. Stoye JP. Koala retrovirus: a genome invasion in real time. Genome Biology. 2006; 7:241. [PubMed: 17118218]
74. Hobbs M, et al. Long-read genome sequence assembly provides insight into ongoing retroviral invasion of the koala germline. Scientific Reports. (under review).
75. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011; 475:493–496. [PubMed: 21753753]
76. Hansen J, Sato M, Russell G, Kharecha P. Climate sensitivity, sea level and atmospheric carbon dioxide. Philosophical Transactions of the Royal Society A. 2013; 371
77. O'Connell JF, Allen J. The process, biotic impact, and global implications of the human colonization of Sahul about 47,000 years ago. Journal of Archaeological Science. 2015; 56:73–84.
78. Clarkson C, et al. Human occupation of northern Australia by 65,000 years ago. Nature. 2017; 547:306–310. [PubMed: 28726833]
79. Saltré F, et al. Climate change not to blame for late Quaternary megafauna extinctions in Australia. Nature Communications. 2016; 7
80. Neaves LE, et al. Phylogeography of the Koala, (*Phascolarctos cinereus*), and harmonising data to inform conservation. PLoS One. 2016; 11:e0162207. [PubMed: 27588685]
81. Dennison S, et al. Population genetics of the koala (*Phascolarctos cinereus*) in north-eastern New South Wales and south-eastern Queensland. Australian Journal of Zoology. 2017
82. Wang J. Triadic IBD coefficients and applications to estimating pairwise relatedness. Genetics Research. 2007; 89:135–153.
83. Wang J. COANCESTRY: a program for simulating, estimating and analysing relatedness and inbreeding coefficients. Molecular Ecology Resources. 2011; 11:141–145. [PubMed: 21429111]
84. Roberts RG, et al. New ages for the last Australian megafauna: continent-wide extinction about 46,000 years ago. Science. 2001; 292:1888–1892. [PubMed: 11397939]
85. Field J, Wroe S, Trueman CN, Garvey J, Wyatt-Spratt S. Looking for the archaeological signature in Australian megafaunal extinctions. Quaternary International. 2013; 285:76–88.
86. Tsangaras K, et al. Historically low mitochondrial DNA diversity in koalas (*Phascolarctos cinereus*). BMC Genetics. 2012; 13:92. [PubMed: 23095716]
87. Taylor A, Graves JM, Murray N, Sherwin W. Conservation genetics of the koala (*Phascolarctos cinereus*) II. Limited variability in minisatellite DNA sequences. Biochemical Genetics. 1991; 29:355–363. [PubMed: 1747097]
88. Taylor AC, et al. Conservation genetics of the koala (*Phascolarctos cinereus*): low mitochondrial DNA variation amongst southern Australian populations. Genetical Research. 1997; 69:25–33. [PubMed: 9164173]
89. Cristescu R, et al. Inbreeding and testicular abnormalities in a bottlenecked population of koalas (*Phascolarctos cinereus*). Wildlife Research. 2009; 36:299–308.
90. Frankham R, et al. Predicting the probability of outbreeding depression. Conservation Biology. 2011; 25:465–475. [PubMed: 21486369]
91. Frankham R, , et al. Genetic Management of Fragmented Animal and Plant Populations. Oxford University Press; 2017.
92. Walker BJ, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS One. 2014; 9:e112963. [PubMed: 25409509]
93. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Research. 1999; 27:573–580. [PubMed: 9862982]
94. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011; 12:491. [PubMed: 22192575]
95. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics. 2012; 13:329–342.
96. Smit A, Hubley R, Green P. RepeatModeler Open-1.0. 2008-2015. 2014.
97. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. 2013–2015. 2015.

98. Boutet E, et al. UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. Plant Bioinformatics: Methods and Protocols. 2016:23–54.

99. O'Leary NA, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research. 2015:733–745.

100. Wong ES, Papenfuss AT, Belov K. Immunome database for marsupials and monotremes. BMC Immunology. 2011; 12:48. [PubMed: 21854560]

101. Grabherr MG, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology. 2011; 29:644–652.

102. Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004; 5:59. [PubMed: 15144565]

103. Borodovsky M, Lomsadze A. Gene identification in prokaryotic genomes, phages, metagenomes, and EST sequences with GeneMarkS suite. Current Protocols in Bioinformatics. 2011:4.5. 1–4.5. 17.

104. Stanke M, et al. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Research. 2006; 34:W435–W439. [PubMed: 16845043]

105. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Research. 2003; 13:2178–2189. [PubMed: 12952885]

106. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution. 2013; 30:772–780. [PubMed: 23329690]

107. Vilella AJ, et al. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Research. 2009; 19:327–335. [PubMed: 19029536]

108. Pond SLK, Muse SV. HyPhy: hypothesis testing using phylogeniesStatistical Methods in Molecular Evolution. Springer; 2005. 125–181.

109. Delport W, Poon AF, Frost SD, Pond SLK. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics. 2010; 26:2455–2457. [PubMed: 20671151]

110. Dobin A, Gingeras TR. Mapping RNA-seq Reads with STAR. Current Protocols in Bioinformatics. 2015:11.14. 1–11.14. 19.

111. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2013; 30:923–930. [PubMed: 24227677]

112. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014; 15:550. [PubMed: 25516281]

113. 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. 2016; 11:e0157022. [PubMed: 27280887]

114. Camacho C, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009; 10:421. [PubMed: 20003500]

115. Löber U, et al. Degradation and remobilization of retroviruses by recombination during the earliest stages of genomic invasion. (In preparation).

116. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26:841–842. [PubMed: 20110278]

117. Sonnhammer EL, Durbin R. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene. 1995; 167:GC1–GC10. [PubMed: 8566757]

118. Phillips SS. Population trends and the koala conservation debate. Conservation Biology. 2000; 14:650–659.

119. Lynch M, et al. Genetic drift, selection and the evolution of the mutation rate. Nature Reviews Genetics. 2016; 17:704–714.

120. 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 Research. 2015; 25:1125–1134. [PubMed: 26129709]

121. Bragg JG, Potter S, Bi K, Moritz C. Exon capture phylogenomics: efficacy across scales of divergence. Molecular Ecology Resources. 2015

122. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012; 9:357–359. [PubMed: 22388286]

123. McKenna A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010; 20:1297–1303. [PubMed: 20644199]

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

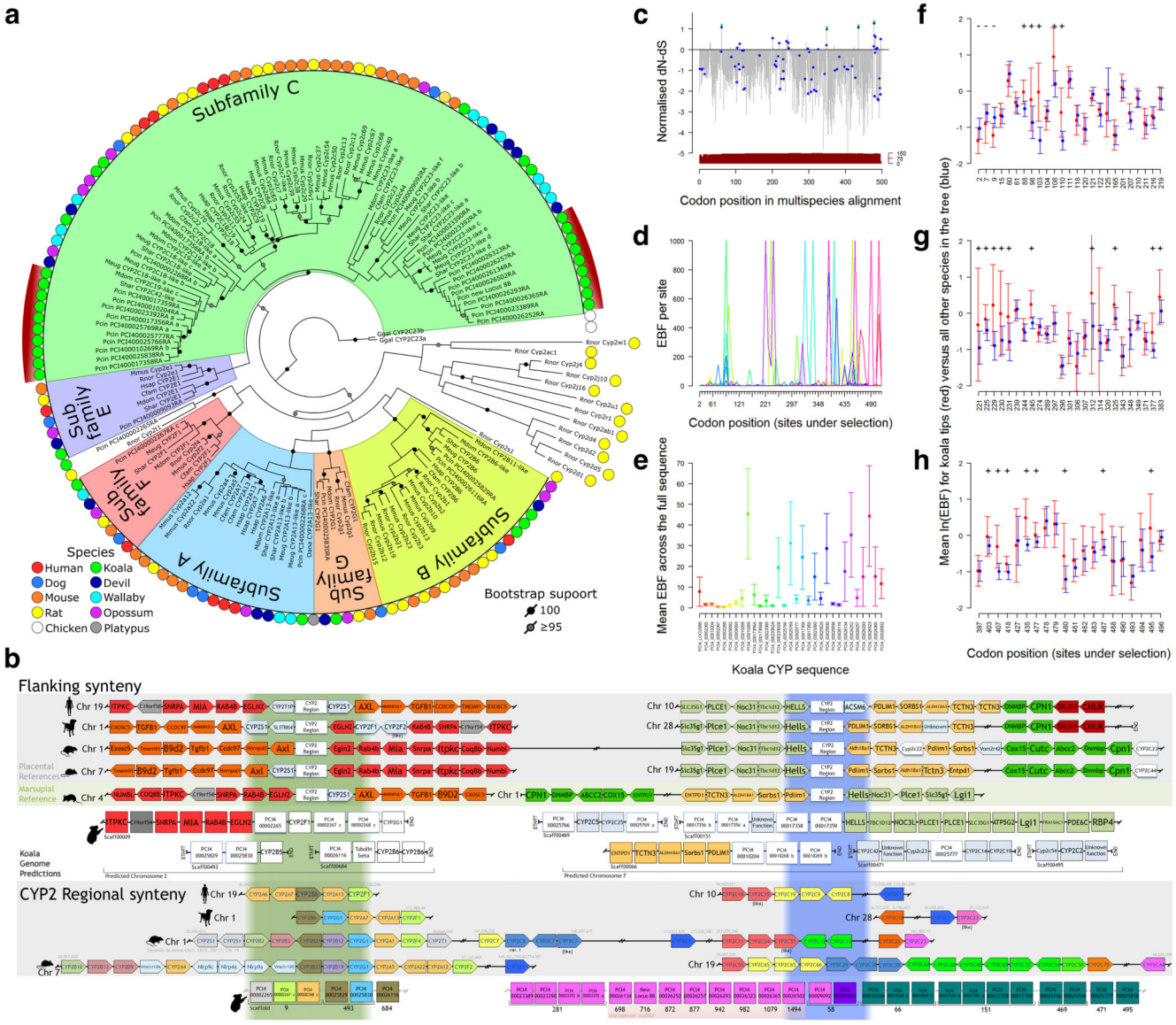

Figure 1. Analysis of Cytochrome P450 family 2 subfamily C gene family expansions, synteny and codons under selection.

a Phylogenetic tree of the *CYP2* gene family in the koala (31 members of CYP2), as compared with marsupials: tammar wallaby, Tasmanian devil, gray short-tailed opossum; and eutherian mammals: human, rat, mouse, dog, platypus; and outgroup chicken. Two independent monophyletic expansions are seen in koala, in the *CYP2C* subfamily (highlighted by red sectors).

b *CYP* synteny map showing expansion of *CYP2C* genes in koala and mouse suggesting that this adaptive characteristic has arisen via tandem duplication.

c-h Selection analysis of *CYP* gene expansion**: c,** Normalised dN-dS (SLAC method) across the alignment of 152 *CYP* sequences (reduced to only those sites with data in koala and at least one other species. Points at the end of bars indicate statistically significant (at threshold $\alpha = 0.1$) evidence for codons under selection across the tree, including four sites showing

positive selection across the entire tree (SLAC method; green circular points), and 70 sites showing episodic selection (MEME method; blue diamonds). **d,** Comparison of episodic selection on particular codons across koala *CYP*s (n = 31 sequences); x-axis shows codons with evidence of statistically significant selection anywhere on the tree (as identified in **c**). **e,** Comparison of mean episodic selection among koala *CYPs* n = 70. Points indicate the mean Empirical Bayes Factor EBF for sites under selection for each sequence; error bars 95% confidence interval. **f, g, h**, Mean EBF (natural log transformed, EBF values of 0 excluded) for koala tree tips (n = 31; red) relative to all others (n = 121 nine species [see Methods]; blue). Points show mean, error bars ± 95% confidence interval, evaluated as 1.96 × SEM (utilising sequence depth as sample size – sample sizes shown in red bars in **a**). Codon positions on x-axis refer to the multispecies alignment used in **a**. Symbols above each point indicate that the mean value for koala site falls outside the 95% CI for all other species (above "+", or below "-", i.e. a two-tailed test at $\alpha = 0.05$). All raw statistics shown (unadjusted for multiple comparisons).

Europe PMC Funders Author Manuscripts

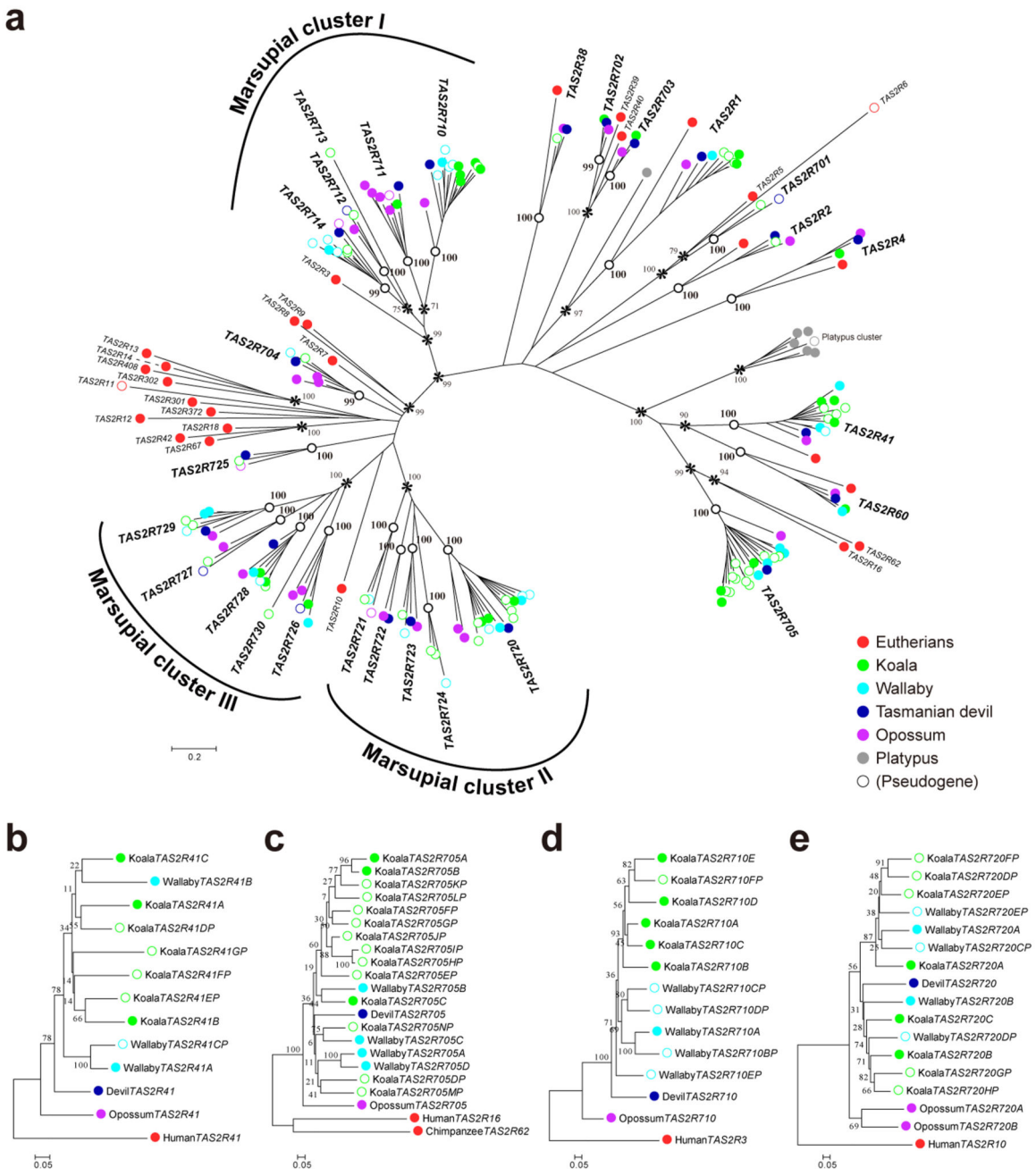

Figure 2. Taste receptor analysis in koalas and other mammals reveals three marsupial specific expansions, and further koala-specific duplications.

TAS2R genes are responsible for bitter taste perception, a role that makes them very important in koalas' need to optimise nutrient content against the high concentration of plant secondary metabolites in the various plants on which they feed. **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 (grey circles) were also used. There were 27

distinct marsupial OGGs (supported by ≥99% bootstrap values), where the nodes of OGG clades were indicated by white open circles Bootstrap values of ≥70% in the nodes connecting OGG clades are also indicated by asterisks. There are three marsupial-specific clusters (named the marsupial cluster I, II and III) where the massive expansion events occurred in the common ancestor of marsupials after split from eutherian ancestors. **b-e**, Reconstructed maximum-likelihood trees of *TAS2R* orthologs in which there are more than 2 duplicates of koala *TAS2R*s were observed; **b** *TAS2R41*, **c** *TAS2R705*, **d** *TAS2R710* and **e** *TAS2R720*. Genomic structure of the umami and sweet taste receptor *TAS1R*s were also analysed and found to be functional in koala (see Supplementary Note 3.7).

Europe PMC Funders Author Manuscripts

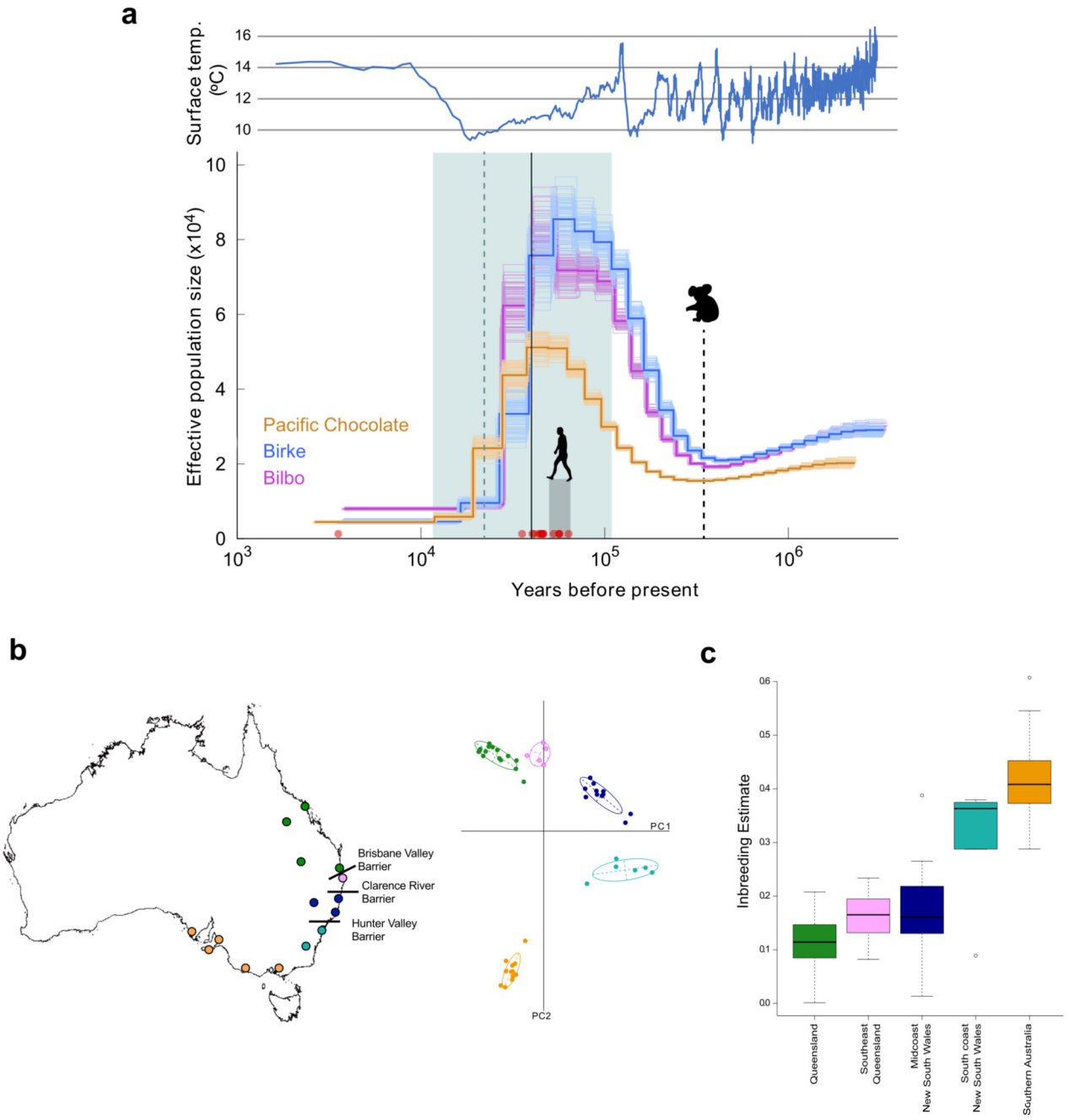

Figure 3. Inference of koala population demographic history and analysis of contemporary koala populations using genome mapped markers.

a Phylogenetic tree of the *CYP2* gene family in the koala (31 members of *CYP2*), as compared with marsupials: tammar wallaby, Tasmanian devil, gray short-tailed opossum; and eutherian mammals: human, rat, mouse, dog, platypus; and outgroup chicken. Two independent monophyletic expansions are seen in koala, in the *CYP2C* subfamily (highlighted by red sectors).

b *CYP* synteny map showing expansion of *CYP2C* genes in koala and mouse suggesting that this adaptive characteristic has arisen via tandem duplication.
c-h Selection analysis of *CYP* gene expansion: **c,** Normalised dN-dS (SLAC method) across 152 *CYP* sequences (sites with data koala and ≥ 1 other species). Points at the end of bars indicate statistically significant ($\alpha = 0.1$) evidence for selection across the tree: four sites show positive selection (SLAC method; green circular points); 70 sites episodic selection (MEME method; blue diamonds). **d,** Episodic selection across koala *CYP*s (n = 31 sequences); x-axis shows codons under selection anywhere on the tree (identified in **c**). **e,** Comparison of mean episodic selection among koala *CYP* sequences (n = 70 codons). **f, g, h**, Mean EBF (natural log transformed, zeroes excluded) for koala tree tips (n = 31; red) relative to all others (n = 121; nine species; blue). Points show mean; error bars ± 95% CI; codon positions as in **a**. A "+" or "-" indicates that the koala mean value falls above/below the 95% CI for all other species (i.e. a two-tailed test at $\alpha = 0.05$). Statistics are unadjusted for multiple comparisons.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

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) *homozygous/**heterozygous	3.42	39.0	1906* 5525** (contigs)	11,589 (contig)	Current study
Platypus (*Ornithorhynchus anatinus)*	2.3	45.5	200,283	959	Warren *et al.* 2008 20
Gray short-tailed Opossum *(Monodelphis domestica)*	3.48	37.7	5,223	59,810	Mikkelsen *et al.* 2007 21
Tammar wallaby *(Notamacropus eugenii)*	2.7	38.8	277,711	37	Renfree *et al.* 2011 22
Tasmanian devil *(Sarcophilus harrisii)*	3.17	36.4	35,974	1,847	Murchison *et al.* 2012 23