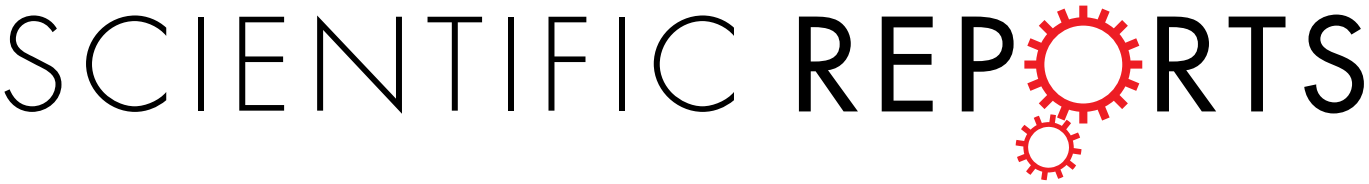

OPEN

The evolutionary history of bears is characterized by gene flow across species

Received: 23 November 2016
Accepted: 17 March 2017
Published: 19 April 2017

Vikas Kumar[1,2], Fritjof Lammers[1,2], Tobias Bidon[1,2], Markus Pfenninger[1,2], Lydia Kolter[3], Maria A. Nilsson[1] & Axel Janke[1,2]

Bears are iconic mammals with a complex evolutionary history. Natural bear hybrids and studies of few nuclear genes indicate that gene flow among bears may be more common than expected and not limited to polar and brown bears. Here we present a genome analysis of the bear family with representatives of all living species. Phylogenomic analyses of 869 mega base pairs divided into 18,621 genome fragments yielded a well-resolved coalescent species tree despite signals for extensive gene flow across species. However, genome analyses using different statistical methods show that gene flow is not limited to closely related species pairs. Strong ancestral gene flow between the Asiatic black bear and the ancestor to polar, brown and American black bear explains uncertainties in reconstructing the bear phylogeny. Gene flow across the bear clade may be mediated by intermediate species such as the geographically wide-spread brown bears leading to large amounts of phylogenetic conflict. Genome-scale analyses lead to a more complete understanding of complex evolutionary processes. Evidence for extensive inter-specific gene flow, found also in other animal species, necessitates shifting the attention from speciation processes achieving genome-wide reproductive isolation to the selective processes that maintain species divergence in the face of gene flow.

Ursine bears are the largest living terrestrial carnivores and have evolved during the last five million years, attaining a wide geographical distribution range (Fig. 1). Bears are a prominent case where conflicting gene trees and an ambiguous fossil record[1] make the interpretation of their evolutionary history difficult[2]. Introgressive gene flow resulting from inter-species mating is believed to be rare among mammals[3]. However, some 600 mammalian hybrids are known[4] and the importance of hybridization has started to gain attention in evolutionary biology[5]. Yet, our knowledge of the extent of post speciation gene flow is limited, because few genomes of closely related species have been sequenced.

In bears, natural mating between grizzlies (brown bears *Ursus arctos*), and polar bears (*Ursus maritimus*) results in hybrid offspring, the grolars[6]. Genome scale studies in brown and polar bears find that 8.8% of individual brown bear genomes have a polar bear origin[7]. Additionally, the brown bear mitochondrial (mt) genome was captured by polar bears during ancient hybridization[8] and polar bear alleles are distributed across brown bear populations all over the world by male-biased migration and gene flow[7,9,10].

Polar and brown bears belong to the sub-family Ursinae, which comprises six extant, morphological and ecological distinct species[11], but hybridization among some ursine bears is possible. A natural hybrid has been reported also between the Asiatic black bear (*Ursus thibetanus*) and the sun bear (*Ursus malayanus*)[12]. In captivity more bear hybrids are known, some of them have been fertile[4]. Despite limited population sizes for most bears and apparently distinct habitats, morphology and ecology, molecular phylogenetic studies have been unable to unequivocally reconstruct the relationship among the six ursine bear species[2]. Especially, the evolution of the American (*Ursus americanus*) and Asiatic black bear is difficult to resolve, despite being geographically separated (Fig. 1).

Evidence from the fossil record, morphology and mitochondrial phylogeny suggested a closer relationship between the Asiatic and the American black bears[13–15]. In contrast, autosomal and Y-chromosomal sequences

[1]Senckenberg Biodiversity and Climate Research Centre, Senckenberg Gesellschaft für Naturforschung, Senckenberganlage 25, D-60325 Frankfurt am Main, Germany. [2]Goethe University Frankfurt, Institute for Ecology, Evolution & Diversity, Biologicum, Max-von-Laue-Str. 13, D-60439 Frankfurt am Main, Germany. [3]AG Zoologischer Garten Cologne, Riehler Straße 173, 50735 Cologne, Germany. Correspondence and requests for materials should be addressed to A.J. (email: Axel.Janke@senckenberg.de)

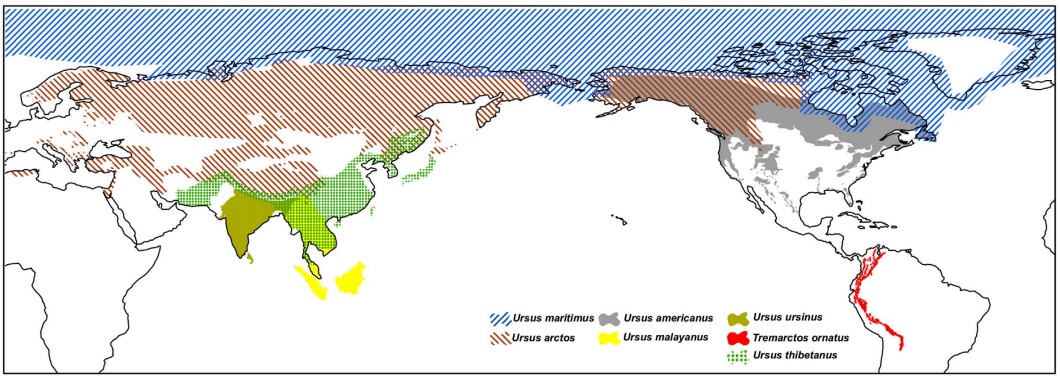

Figure 1. Approximate geographic distribution of extant bears according to IUCN data. Figure has been created using ArcGIS 10 (http://desktop.arcgis.com/en/arcmap/) with base map from GADM v 2.0. (http://www.gadm.org). Species range maps IUCN2015 (http://www.iucnredlist.org).

support a grouping with the American black bear being sister group to the brown/polar bear clade[2,9,16]. Another conflict between mitogenomics, morphology and autosomal sequence data is the position of the morphologically distinct sloth bears (*Ursus ursinus*). Mitochondrial DNA (mtDNA) analyses and morphological studies placed sloth bears outside of all other ursine bears, while nuclear gene analyses favor a position close to sun bears[2,15,17]. A study of nuclear introns with multiple individuals for each ursine species was unable to reconstruct a well-supported species tree and suggested that incomplete lineage sorting (ILS) and/or gene flow caused the complexities in the ursine tree[2]. However, previous molecular studies did not have access to genome data from all bear species and were thus limited to single loci.

The genomic era allows a detailed analyses of how gene flow from hybridization affects genomes, and has revealed much more complex evolutionary histories than previously anticipated for many species, including our own[18–20]. Multiple genomic studies on polar, brown bears and the giant panda[10,21–23] lead to a wealth of available genomic data in these species. We investigated all living Ursinae and Tremarctinae bear species based on six newly sequenced bear genomes and published ones. Methods specifically developed to deal with complex genome data[24,25] and gene flow[18,26] are applied to resolve and understand the processes that have shaped the evolution of bears.

Results

The sequenced individuals were morphologically typical for the respective species. Mapping Illumina reads against the polar bear genome[23] yielded an average coverage of 11X. Supplementary Tables 1 and 2 detail the sequencing and assembly data, and provide accession numbers of the included species. As a basis for subsequent analyses, non-overlapping 100 kb Genome Fragments (GFs) were extracted from polar bear scaffolds > 1 megabase (Mb). These have presumably a higher assembly quality than smaller fragments and still represent > 96% of the genome (Supplementary Fig. 1). Heterozygous sites, gaps, repetitive sequences, and transposable element sequences were removed from GF alignments (Supplementary Fig. 2). Pedigrees (Supplementary Fig. 3) and genome-wide heterozygosity plots (Supplementary Fig. 4) show that the sequenced individuals are neither hybrids nor, compared to wild specimens, severely inbred.

Network analysis depicts hidden conflict in the coalescent species tree. GFs larger than 25 kb, representing the majority of the length distribution (Supplementary Fig. 2), contain on average 104 substitutions among Asiatic bears (Supplementary Fig. 5). Phylogenetic topology testing on real and simulated sequence data shows that GFs with this information content significantly reject alternative topologies (Supplementary Figs 6 and 7). For subsequent coalescence, consensus, and network analyses, only GFs > 25 kb were used and the results are thus based on firmly supported Maximum Likelihood (ML) analyses.

A coalescent species tree utilizing 18,621 GFs > 25 kb (869,313,834 bp) resolved the relationships among bears with significant support for all branches (Fig. 2A, Supplementary Fig. 8). In the coalescent-based species tree, sun and sloth bears are sister group to the Asiatic black bear, and the American black bear groups with polar and brown bears. The spectacled bear is, consistent with previous results[2,16], placed as sister taxon to Ursinae. The well-resolved coalescent species tree appears to be without conflict from genomic data.

However, a network analysis[27] gained from the same 18,621 GFs identifies conflicting phylogenetic signal (Fig. 2B). The square and cuboid-like structures indicate alternative phylogenetic signals, particularly among brown and polar bears, but also among the Asiatic bears. The brown bear from the Admiralty, Baranof, and Chichagof (ABC) islands groups in different arrangements with other brown and polar bears, consistent with gene flow between the two species[7,8,23]. When the threshold level for depicting conflicting branches is reduced in the network analysis, the signal becomes increasingly complex, illustrating the conflict among 18,621 ML-trees (Supplementary Fig. 9). Still, the network analysis agrees with the species tree when the spectacled bear is the outgroup. The phylogenetic conflict can be caused by incomplete lineage sorting (ILS) or gene flow, but less likely from lack of resolution due to the strong phylogenetic signal of each GF (Supplementary Figs 6 and 7). Analyses of 8,050 protein coding sequences (10,303,323 bp) and GFs from scaffolds previously identified as X

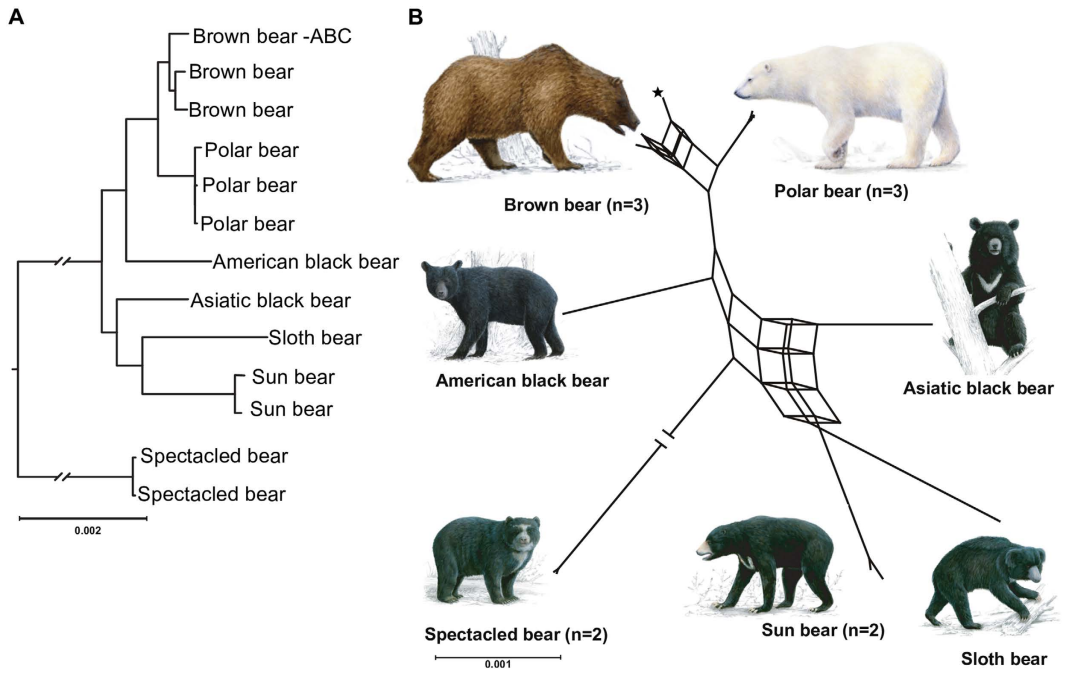

Figure 2. A coalescent species tree and a split network analysis from 18,621 GF ML trees. (**A**) In the coalescent species tree all branches receive 100% bootstrap support. The position of root and depicted branch lengths were calculated from coding sequence and 10 Mb of GF data respectively. (**B**) A split network with a 7% threshold level depicts the complex phylogenetic signal in bear genomes. As expected, the ABC-island brown bear (asterisk) shares alleles with polar bears; among other bears allele sharing is complex. Paintings by Jon Baldur Hlidberg (www.fauna.is).

chromosomal (total 74 Mb)[22], conform to the species tree and networks (Supplementary Fig. 10). Finally, the paternal side of bear evolution based on Y chromosome sequences[28] for available genomes is consistent with the inferred species tree (Supplementary Fig. 11).

The Bayesian mtDNA tree (Fig. 3, Supplementary Fig. 12) conforms to previous studies[2,15], making this the hitherto largest taxonomic sampling of 38 complete bear mt genomes. However, several nodes of the mtDNA tree differ notably from the coalescent species tree (Fig. 2A). In the mtDNA tree, the brown bears are paraphyletic, because the brown bear mt genome introgressed into the polar bear population[8]. The extinct cave bear (*Ursus spelaeus*) is the sister group to polar and brown bears. The American black bear is the sister group to the Asiatic black bear, and the sloth bear is the sister group to all ursine bears. The topological agreement of the mtDNA tree to previous studies and placement of the new individuals corroborates that the studied individuals are representative for their species.

Finally, a consensus analysis based on GF ML-trees (Supplementary Fig. 13) produces a tree that is identical to the coalescent species tree, but highlights that numerous individual GF trees support alternative topologies (Supplementary Table 3). Inspection of the individual 18,621 GF ML topologies shows that 38.1% (7,086) support a topology where Asiatic black bear is the sister group to the American black/brown/polar bear clade. The Asiatic black bear groups in different arrangements with the two other Asiatic bears: 18.7% (3,474) of the branches support a grouping with the sun bear, and 7.5% (1,394) with the sloth bear.

Gene flow among bears is common. Seemingly conflicting phylogenetic signals in evolutionary analyses can be explained by incomplete lineage sorting (ILS) or gene flow among species. In contrast to the largely random process of ILS, gene flow produces a bias in the phylogenetic signal, because it is a directed process. The *D*-statistic measures the excess of shared polymorphisms of two closely related lineages with respect to a third lineage[18] and can thus discriminate between gene flow and ILS. The test assumes that the ancestral population of the in-group taxa was randomly mating and recently diverged[29]. These assumptions might be compromised in wide-spread, structured species like bears. However, speciation is rarely instantaneous, but is rather preceded by a period of population divergence. This should not compromise the test as long as there was a panmictic population ancestral to the progenitor populations of the eventual daughter species at some point in time, which is a reasonable assumption.

The *D*-statistics analyses find evidence of gene flow between most sister bear species (Fig. 4, Supplementary Tables 4 and 5 and Supplementary Fig. 14). Regardless if spectacled bear or giant panda is used as outgroup, the involved species and relative signal strengths of gene flow in the tested topologies remain the same (Supplementary Table 6). The *D*-statistics is limited to four-taxon topologies and therefore gene flow signals are difficult to interpret when they occur between distant species, as it cannot determine if it is a direct, indirect, or ancestral signal. For taking more complex gene flow patterns into account, and to determine the direction of gene

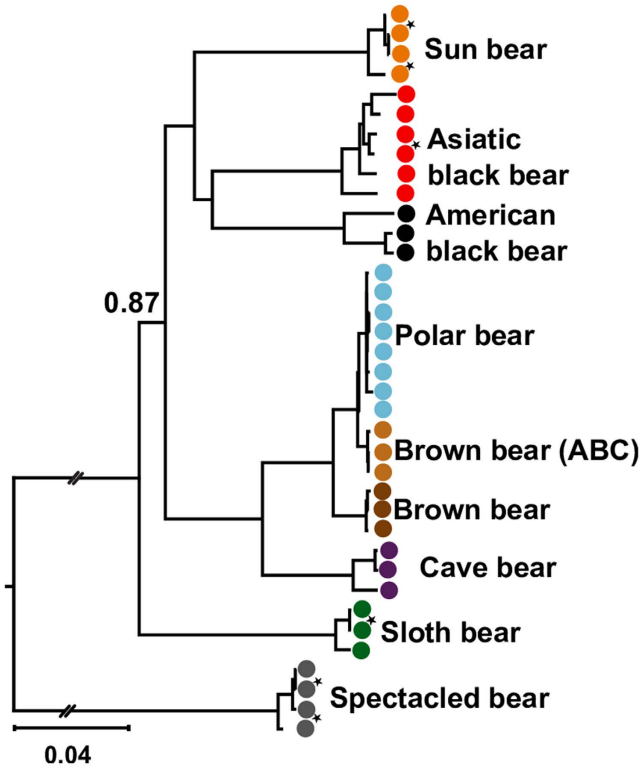

Figure 3. Phylogenetic relationship among the bears using mtDNA genomes. A Bayesian tree from 37 complete mt genomes (colored circles) and stars indicate the new mt genomes. The tree is rooted with panda genome (not shown). Supplementary Fig. 12 shows support values for $p < 1.0$ and accession numbers.

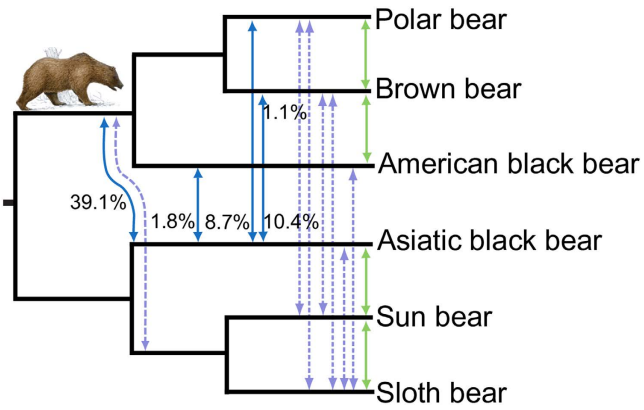

Figure 4. Graphical summary of gene flow analyses using *D* and D_{FOIL} statistics on a cladogram. D_{FOIL} analyses estimated the percentage of GFs rejecting the species tree and indicating gene flow. Blue arrows show values >1%, and dashed lavender for <0.1% (Table 1). These percentages do not indicate the amount of introgressed DNA, which can be a fraction of the GF sequence. Green arrows depict significant *D*-statistics data for gene flow signal. Some gene flow cannot have occurred directly between species, because the species exist in different habitats, but may be remnants of ancestral gene flow or gene flow through a vector species.

flow, we applied the recently introduced D_{FOIL}-statistics[26]. This method uses a symmetric five-taxon topology and has specifically been developed to detect and differentiate gene flow signal among ancestral lineages[26].

In agreement with the phylogenetic conflict and *D*-statistics, the D_{FOIL}- statistics finds gene flow between the ancestor of the American black bear/brown/polar bear clade and the Asiatic black bear (Fig. 4, Table 1). The Etruscan bear was geographically overlapping with other bear species and was, like the Asiatic black bear, widely distributed[30]. It has been identified in fossil layers of Europe 2.5 Ma − 1.0 Ma[1,31]. The wide geographical distribution would explain the nearly equally strong gene flow from Asiatic black bear into brown bear also observed in the *D*-statistics (Supplementary Fig. 14). Finally, there is a gene flow signal between the American and Asiatic black bears. The gene flow could have taken place either on the American or Asiatic side of the Bering Strait and

	AmB, BrB, SuB, AsB	AmB, BrB, SlB, AsB	AmB, PoB, SuB, AsB	AmB, PoB, SlB, AsB	Average % (As/AmB, SuB/SlB, PoB/BrB)	Average % (BrB, AmB, SuB/SlB, AsB)	Average % (PoB, AmB, SuB/SlB, AsB)
AsB = > AmB	0.07% (15)	0.08% (18)	0.14% (31)	0.15% (34)	0.11%		
AmB = > AsB	1.22% (270)	1.42%(313)	2.06% (454)	2.48% (547)	1.80%		
Su/SlB = > AmB	0.02% (4)	0.01% (1)	0.02% (5)	0.02% (5)		0.01%	0.02%
AmB = > SuB/SlB	0.02% (4)	0.01%(1)	0.02% (4)	0.01% (2)		0.01%	0.01%
SuB/SlB = > BrB/PoB	0.07% (16)	0.02%(4)	0.02% (5)	0.02% (5)		0.05%	0.02%
BrB/PoB = > SuB/SlB	0.03% (6)	0.01%(1)	0.02% (5)	0.01% (3)		0.02%	0.02%
AsB = > BrB/PoB	1.25% (276)	1.02%(225)	0.46% (101)	0.29 (64)		1.14%	0.37%
Br/Po = > As	9.70% (2159)	11.0% (2415)	8.37% (1846)	9.02 (1989)		10.4%	8.69%
BrB/PoB, AmB < = > SuB/SlB	0.10% (23)	0.06 (14)	0.20% (44)	0.11% (25)	0.12%		
BrB/PoB, AmB < = > AsB	32.2% (7098)	32.0% (7060)	46.3% (10214)	45.8% (10108)	39.1%		

Table 1. Gene flow detected by the D_{FOIL} analyses that is based on a five taxon analysis. The table shows the percentage of 100 kb fragments that have a signal of gene flow, and in brackets the absolute number is shown. The rows show these values for different combinations of four bear species with the spectacled bear as an outgroup. The last three columns summarize amount of gene flow. The arrows in the table (= >) indicate the direction of the gene flow, between the respective species for each of the combinations analyzed. For example: between Asiatic black and American black bear the D_{FOIL} finds 15–34 GF that support gene flow (first row). There is much more gene flow in the other direction (second row). Abbreviations: SuB (Sun bear), SlB (Sloth bear), AsB (Asiatic black bear), AmB (American black bear), BrB (Brown bear, Finland), and PoB (Polar bear, Svalbard).

is consistent with mitochondrial capture between the species[2] (Fig. 3). Most of the weaker gene flow signals in Fig. 4 (dashed-lines) do not necessarily reflect direct species hybridization and are possibly remnants of ancestral gene flow not detected due to allelic loss or signals of indirect gene flow by ghost lineages or intermediate species. Permutations of species for the D_{FOIL} analysis including other polar, sloth and brown bear individuals show that the results are taxon independent (Table 1).

PhyloNet[32] has been developed to detect hybridization events in genomic data while accounting for ILS. We applied the ML approach implemented in PhyloNet[32] to detect hybridization among bear species. Due to computational constraints we sampled 4,000 ML trees from putatively independent GFs using one individual representing per species. The ABC island brown bear was chosen as another representative for brown bears and positive control, because its population hybridized with polar bears[7,8,28]. The outgroup, the spectacled bears were removed to reduce the computational complexity and, because previous analyses using *D*-statistics and D_{FOIL} did not detect gene flow between tremarctine and ursine bears. The complex phylogeny requires exceptional computational time so we analyzed only networks with up to two reticulations. The resulting PhyloNet network with the highest likelihood (Supplementary Fig. 15) shows reticulations between ABC island brown bear and polar bears, and also between the Asiatic black bear and the ancestral branch to American black, brown and polar bears. It is noteworthy, that the second reticulation has a high inheritance probability (41.8%), which agrees with the strongest gene flow signal identified by D_{FOIL} analyses (Fig. 4, Table 1). Due to computational limits so far only two reticulations that represent the strongest hybridization signals were identified. For three and more reticulations the network-space becomes extremely large.

Additional analysis using CoalHMM[33] supports the findings of gene flow from *D*-, D_{FOIL}, and PhyloNet analyses (Supplementary Fig. 16). It shows that a migration model fits most pair wise comparisons significantly better than ILS, and is robust under a broad range of parameters (Supplementary Figs 17 and 18). Thus, gene flow among bears throughout most of their history is the major factor for generating conflicting evolutionary signals.

Estimation of divergence times and population splits. The phylogenomic divergence time estimates (Fig. 5) are older than previous estimates based on nuclear gene data[2], but consistent with that from mtDNA data[15] (Supplementary Table 7). The amount of heterozygous sites differs among species and individuals, and is highest in the Asiatic black bear genome and, as expected[2] lowest in the polar bears and spectacled bears (Supplementary Fig. 4). It is noteworthy that the average numbers of heterozygous sites differ among the two sun bears, which may reflect different population histories.

Estimates for past changes in effective population size (N_e) using the pairwise sequentially Markovian coalescent (PSMC)[34] are shown in Fig. 6 (Supplementary Fig. 19). While PSMC plots from low coverage genomes may vary and not be ultimately accurate, the plots inferred for the brown, polar and American black bear are very similar to previous published on higher coverage genome (Supplementary Fig. 20)[10]. The demographic histories of the Asian bear individuals vary widely, but do not overlap in bootstrap analyses since 100 ka (Supplementary Fig. 21).

Discussion

Previously, nuclear gene trees and mitochondrial trees have been in conflict[14–16], and a forest of gene trees made it difficult to conclusively reconstruct the relationships among bears, in particular among Asiatic bears[2]. Now,

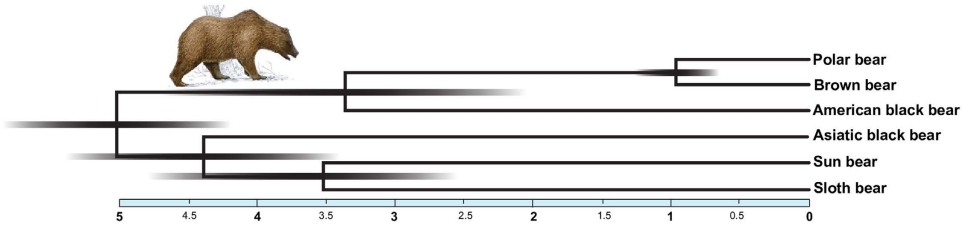

Figure 5. Phylogenomic estimates of divergence times. The scale bar shows divergence times in million years and 95% confidence intervals for divergence times are shown as shadings (Supplementary Table 7). The tree is rooted with the panda genome (not shown).

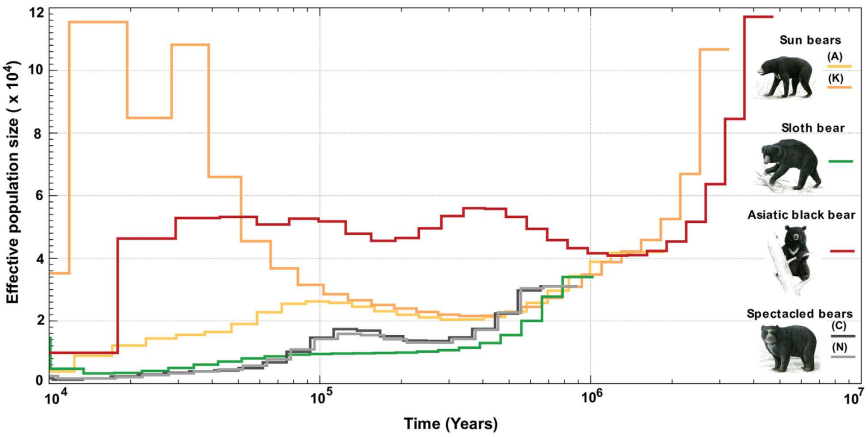

Figure 6. Historical effective population sizes (N_e) using the pairwise Markovian coalescent (PSMC) analyses for the newly sequenced bear genomes. X-axis:time, y-axis:effective population size (N_e). The two sun bears have radically different, non-overlapping population histories (Supplementary Fig. 21). The Asiatic black bear had a constant large N_e since 500 ka similar to that of the brown bear and consistent with a wide geographic distribution and high heterozygosity (Supplementary Fig. 4).

phylogenomic analyses resolve a solid coalescent species tree and provide a temporal frame of the evolutionary history of the charismatic ursine and tremarctine bears and allow a glimpse into their demographic history.

According to the PSMC analyses the Asiatic black bear maintained a stable and a relatively high long-term *Ne* since 500 ka (Fig. 6). This is consistent with its wide geographic distribution and its high degree of heterozygous sites in the genome[2]. The effective population size of the Asiatic black bear declined some 20 ka, correlating with the end of the later part of the ice age. By contrast, the spectacled bear maintained a relatively low long-term effective population size, consistent with their lower population diversity[2,35]. The demography of two sun bear individuals is strikingly different from each other since 100 ka. As the bootstrap replicates do not overlap, the different curves support a hypothesis of separate population dynamics (Supplementary Fig. 21). Their distinct mitochondrial lineages (Fig. 3) might indicate that the two sun bear individuals belong to the described subspecies *U. m. malayanus* (Sumatra and Asian mainland) and *U. m. euryspilus* (Borneo) respectively[36]. The ancestor of extant sun bears might have settled in the Malay Archipelago during the marine isotope stage (MIS)[6]. In the following Eemian interglacial, Borneo got isolated, thereby giving rise to different environmental conditions and to a distinct sun bear subspecies, but without samples from multiple individuals from known locations and high coverage genomes, this remains speculative.

Multi-species-coalescent methods that are becoming increasingly important in genomic analyses[37] taking phylogenetic conflict into account. However, when analyzing GFs > 25 kb, phylogenetic conflict is not caused by noise, but by evolutionary signal and should not be ignored[38]. Phylogenetic networks show that evolutionary histories of numerous GFs, i.e. various regions of their genome, are significantly different, not only because the phylogenetic signal differs drastically, but it does so with statistically significant support. This is also evident from large-scale evolutionary analysis of insertion patterns of transposable elements into the bear genomes, which yield a similarly complex history of bears[39]. Compared to a study based on 14 loci[2] we were able to fully resolve the species relationship among Ursidae. In addition genome analyses shows that, the conflicting relationship shown in[2] are to be the result of gene flow which is not only limited to sister species. It is important to realize that bifurcating species trees, even coalescence based, can only convey a fraction of the evolutionary information contained in entire genomes and that network analyses are needed to identify underlying conflict in the data[24,38]. The analyses of the ursine phylogeny suggest that gene flow and not incomplete lineage sorting are major cause for the reticulations in the evolutionary tree. These two processes can be distinguished from each other by methods and programs like *D*-statistics, D_{FOIL} and Phylo-Net[18,26,32] that are specifically developed for this task.

Some of the inferred gene flow between bear species appears weak or episodic and thus requires further corroboration by additional sampling of individuals. Population analyses show that American black bears are divided into two distinct clades that diverged long before the last glacial maximum, indicating a long and isolated evolutionary history on the North American continent[40]. Thus, it is unlikely that American black bears came into contact with the Asiatic sun and sloth bears[40]. Likewise, introgressive gene flow between south-east Asiatic bear species and polar bears requires an explanation, because they have been evolving in geographically and climatically distinct areas, from the time when polar bears diverged from brown bears and began parapatric speciation in the Arctic. It is therefore possible that some gene flow events occurred through an intermediate species. The brown bear has been shown to distribute polar bear alleles across its range[7] and may therefore be a plausible vector species for genetic exchange between Asiatic bears and the polar, or American black bear. The brown bear is a likely extant candidate, because it has been and is geographically wide-spread[41]. Furthermore, the geographical range of brown bears overlaps with all other ursine bear species (Fig. 1), they have reportedly migrated several times across continents and islands[41], and numerous brown bear hybrids with other bears in either direction are known[4]. While also the Asiatic black bear was widely distributed across Asia and had, like the brown bear[10], a large effective population size (Fig. 6), a migration of the Asiatic black bear into North America has not been shown. Likewise, migration of the American black bear in the opposite direction, from the American to the Asian continent, is not evident from fossil data. The D_{FOIL} and PhyloNet analyses[26,32] are powerful tools to detect ancestral gene flow, such as the prominent signal between the Asiatic black bear and the ancestor to the American black, brown and polar bears (Fig. 4, Table 1). In fact, gene flow during early ursine radiation from extinct bear species, such as the Etruscan bear or the cave bear is to be expected to leave signatures in genomes of their descendants and thus causing conflict in a bifurcating model of evolution.

Speciation as a selective rather than an isolation process. There is no question that bears are morphologically, geographically and ecologically distinct and they are unequivocally accepted as species even by different species concepts[42]. Yet, our genome-wide analyses identify gene flow among most ursines, making their genome a complex mosaic of evolutionary histories. Increasing evidence for post-speciation gene flow among primates, canines, and equids[19,20] suggests that interspecific gene flow is a common biological phenomenon. The occurrences of gene flow and to a lesser extent ILS, of which a fraction in the phylogenetic signal cannot be excluded, suggest that the expectation of a fully resolved bifurcating tree for most species might be defied by the complex reality of genome evolution. Recent genome-scale analyses of basal divergences of the avian[43], and even metazoan[44] tree share the same difficulties to resolve certain branches as observed for mammals[45]. Detecting gene flow for these deep divergences is difficult and therefore most of the reticulations and inconsistent trees have so far been attributed to ILS[46].

The recent discoveries of gene flow by introgressive hybridization in several mammalian species[19,20] and in bears over extended periods of their evolutionary history have a profound impact of our understanding of speciation. If, in fact gene flow across is frequent, and can last for several hundred-thousand years after divergence, evolutionary histories of genomes will be inherently complex and phylogenetic incongruence will depict this complexity. Therefore, speciation should not only be viewed as achieving genome-wide reproductive isolation but rather as selective processes that maintain species divergence even under gene flow[47].

Materials and Methods

Genome sequencing, mapping and creation of consensus sequences. Prior to sampling and DNA extraction and evolutionary analyses, pedigrees from zoo studbooks and appearance of the individuals confirmed that these individuals are not hybrids (Supplementary Fig. 3). DNA extraction from blood samples was done in a pre-PCR environment on different occasions to avoid confusion by standard phenol/chloroform protocols and yielded between 1 to 6 μg DNA for each of the six bear individuals (Supplementary Table 1). Paired end libraries (500 bp) were made by Beijing Genome Institute (BGI) using Illumina TrueSeq and sequencing was done on Illumina HiSeq2000 resulting in 100 bp reads. Routine diagnosis samples were taken by a veterinarian and stored for later analyses in accordance with ethical guidelines of the respective institutions (see Acknowledgements), were used opportunistically for DNA isolation in accordance to best ethical and experimental practice of the Senckenberg Natural Research Society.

Raw reads were quality-trimmed by Trimmomatic[48] with a sliding window option, minimum base quality of 20 and minimum read length of 25 bp. The assembled polar bear genome[23] was used for reference mapping using BWA version 0.7.5a[49] with the BWA-MEM algorithm on scaffolds larger than 1 Mb. Scaffolds shorter than 1 Mb in length were not involved in the mapping and analyses, due to potential assembly artefacts[50] and for reducing the computational time in downstream analyses. Duplicate Illumina reads were marked by Picard tools version 1.106 (http://picard.sourceforge.net/) and the genome coverage was estimated from Samtools version: 0.1.18[51].

Freebayes version 0.9.14–17[52] called Single Nucleotide Variants (SNVs) using the option of reporting the monomorphic sites with additional parameters as -min-mapping-quality 20, -min-alternate-count 4, -min-alternate-fraction 0.3 and -min-coverage 4 with insertion/deletion (indel) realignment. A custom Perl script created consensus sequences for each of the mapped bear individuals from the Variant Call Format (VCF) files, keeping the heterozygous sites and removing indels. In order to complete the taxon sampling of the ursine bears, reads from six previously published genomes (Supplementary Table 1) selected and on the basis of geographic distribution, availability and sequence depth and SNVs were called as described above. For the two high coverage (>30X) genomes, SNVs calling parameters (-min-coverage) were set as one-half of the average read depth after marking duplicates. Genome error rates[53] were calculated on the largest scaffold (67 Mb) for all bear genomes, confirming a high quality of the consensus sequences. (Supplementary Methods and Supplementary Fig. 22).

Data filtration, simulation of sequence length and topology testing. The next step was to create multi-species alignments for further phylogenetic analysis from all 13 bear individuals. In order to create a data set with reduced assembly and mapping artefacts, genome data was masked for TEs and simple repeats[19] using the RepeatMasker[54] output file of the polar bear reference genome available from http://gigadb.org/[23]. Since the polar bear reference genome RepeatMasker output file did not contain the simple repeat annotation, we repeat-masked the polar bear reference genome with the option (-int) to mask simple repeats. Next all bear genomes were masked with bedtools version 2.17.0[55] and custom Perl scripts. Non-overlapping, sliding window fragments of 100 kb were extracted using custom perl scripts together with the program splitter from the Emboss package[56] (Supplementary Fig. 1), creating a dataset of 22,269 GFs from 13 bear individuals. Heterozygous sites, and repeat elements were all marked "N" and removed using custom Perl scripts. An evaluation of the minimum sequence length of GFs needed for phylogenetic analysis was done by estimating how much sequence data is needed to reject a phylogenetic tree topology using the approximate unbiased, AU test[57]. Only sufficiently long sequences can differentiate between alternative trees with statistical significance. The evaluation was done in two separate analyses: (a) with a simulated data set and (b) on a data set of 500 random GFs (Supplementary Methods).

Phylogenetic analysis using Genomic Fragment (GF), coding and mitochondrial sequences. For phylogenetic analysis, all GFs with length < 25 kb were removed from the initial 22,269 GFs resulting in a data set consisting of 18,621 GFs (mean sequence length of 46,685 bp and standard deviation of 9,490 bp). The dataset was then used to create a coalescent phylogenetic species tree. First the selected GFs were used to create individual ML-trees using RAxML version 8.2.4[58]. The best fitting substitution model was selected on 10 Mb of genomic data using jModelTest 2.1.1[59] available in RAxML version 8.2.4[58] and applied to all ML analyses. From 18,621 ML trees, ASTRAL[25] constructed a coalescent species tree. For bootstrap support of the coalescent species tree, GF ML trees were bootstrapped 100 times, generating a total of 1,862,100 ML trees. The bootstrapped ML-trees and the coalescent species tree were used as input in ASTRAL[25] using default parameters to generate bootstrap support. The consense program in Phylip version 3.69[60] built from 18,621 ML-trees, a majority rule consensus tree. SplitsTree version 4[61] created a consensus network from the 18,621 GF ML-trees with various threshold settings (5%, 7%, 10% and 30%), to explore the phylogenetic conflict among the bear species. Similarly phylogenetic analysis of nuclear protein coding sequences (CDS) and mitochondrial genomes were done with panda genome as outgroup (Supplementary methods).

Gene flow analysis using *D*-statistics and the D_{FOIL}-method. The program ANGSD[62] was used for admixture analysis (*D*-statistics) among the ursine bears using the spectacled bear-Chappari as outgroup. The reads of the other bears were mapped to the consensus sequence of the spectacled bear as described in method section. In addition, indel realignment was done using GATK version 3.1–1[63]. All possible four-taxon topologies of the bear species including sun bear-Anabell, brown bear-Finland, Brown bear-ABC, Polar bear-2, American black bear, Asiatic black bear, Sloth bear were involved for gene flow analysis using *D*-statistics. A block jackknife procedure (with 10 Mb blocks) with parameters: -minQ 30 and -minMapQ30, was used to assess the significance of the deviation from zero. We also mapped the sun bear-Anabell, the Asiatic black bear and the sloth bear against the giant panda genome (ailMel1) http://hgdownload.soe.ucsc.edu/goldenPath/ailMel1/bigZips/ and repeated the analyses described above on to investigate if the outgroup choice affected our conclusions. In addition, we analyzed the data using D_{FOIL}-statistics[26], to detect signatures of introgression. For this analysis we assumed the coalescent species tree (Fig. 2A) and selected a window size of 100 kb with–mode dfoil as suggested by the authors[26]. Other parameters were left at default.

Hybridization inference using PhyloNet. A data set of 4,000 random (every fourth) GFs, that are putatively in linkage equilibrium, was created to calculate rooted ML trees with RAxML as described earlier. The trees were pruned to contain one individual of each ursine species plus the ABC- brown bear to reduce computational complexity of the ML analyses. Maximum likelihood networks in a coalescent framework, thus incorporating ILS and gene flow, were inferred using PhyloNet[32,64] allowing 0, 1 and 2 reticulations in 50 runs and returning the five best networks.

Estimation of heterozygosity, past effective population size and divergence times. In order to calculate the amount of heterozygous sites as well as their distribution in all the bear genomes, their genomes were fragmented into 10 Mb regions using custom Perl scripts. The number of heterozygous sites was counted using a custom Perl script and plotted as distributions using R. The pairwise sequentially Markovian coalescent (PSMC)[34] analysis assessed past changes in effective population size over time. We used default parameters and 100 bootstrap replicates assuming a generation time for brown and polar bears of ten years, and six years for the other bear species for the PSMC analysis. We selected a mutation rate of 1×10^{-8} changes/site/generation for all species. These parameters were used in previous brown and polar bear analyses[10] and enable comparability between the studies. A generation time of six years has been shown for the American black bear[65] and was deemed realistic for the other relatively small-bodied bears. The mutation rate is close to a pedigree-based mutation rate of 1.1×10^{-8} changes/site/generation in humans[66] that is considered to be typical for mammals. We also estimated the divergence time for all the bear species (Supplementary methods).

References

1. Wagner, J. Pliocene to early Middle Pleistocene ursine bears in Europe: a taxonomic overview. *J. Natl. Mus. Prague Nat. Hist. Ser.* **179,** 197–215 (2010).
2. Kutschera, V. E. *et al.* Bears in a Forest of Gene Trees: Phylogenetic Inference Is Complicated by Incomplete Lineage Sorting and Gene Flow. *Mol. Biol. Evol.* **31,** 2004–2017 (2014).
3. Coyne, J. A. & Orr, H. A. *Speciation*. **37**, (Sunderland, MA: Sinauer Associates, 2004).

4. Gray, A. *Mammalian hybrids. A check-list with bibliography*. (Commonwealth Agricultural Bureaux, 1972).
5. Mallet, J. Hybridization as an invasion of the genome. *Trends Ecol. Evol.* **20,** 229–237 (2005).
6. Smol, J. P. Climate Change: A planet in flux. *Nature* **483,** S12–S15 (2012).
7. Cahill, J. A. *et al.* Genomic evidence of geographically widespread effect of gene flow from polar bears into brown bears. *Mol. Ecol.* **24,** 1205–1217 (2015).
8. Hailer, F. *et al.* Nuclear genomic sequences reveal that polar bears are an old and distinct bear lineage. *Science* **336,** 344–347 (2012).
9. Bidon, T. *et al.* Brown and polar bear Y chromosomes reveal extensive male-biased gene flow within brother lineages. *Mol. Biol. Evol.* **31,** 1353–1363 (2014).
10. Miller, W. *et al.* Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. *Proc. Natl. Acad. Sci.* **109,** E2382–E2390 (2012).
11. Nowak, R. *Walker's Mammals of the World*. (Johns Hopkins Press, 1991).
12. Galbreath, G. J., Hunt, M., Clements, T. & Waits, L. P. An apparent hybrid wild bear from Cambodia. *Ursus* **19,** 85–86 (2008).
13. McLellan, B. & Reiner, D. A review of bear evolution. *Bears Their Biol. Manag.* 85–96 (1994).
14. Yu, L., Li, Y. W., Ryder, O. A. & Zhang, Y. P. Analysis of complete mitochondrial genome sequences increases phylogenetic resolution of bears (Ursidae), a mammalian family that experienced rapid speciation. *BMC Evol. Biol.* **7,** 198 (2007).
15. Krause, J. *et al.* Mitochondrial genomes reveal an explosive radiation of extinct and extant bears near the Miocene-Pliocene boundary. *BMC Evol. Biol.* **8,** 220 (2008).
16. Pagès, M. *et al.* Combined analysis of fourteen nuclear genes refines the Ursidae phylogeny. *Mol. Phylogenet. Evol.* **47,** 73–83 (2008).
17. Abella, J. *et al.* Kretzoiarctos gen. nov., the Oldest Member of the Giant Panda Clade. *PLoS ONE* **7,** e48985 (2012).
18. Green, R. E. *et al.* A Draft Sequence of the Neandertal Genome. *Science* **328,** 710–722 (2010).
19. Carbone, L. *et al.* Gibbon genome and the fast karyotype evolution of small apes. *Nature* **513,** 195–201 (2014).
20. Jónsson, H. *et al.* Speciation with gene flow in equids despite extensive chromosomal plasticity. *Proc. Natl. Acad. Sci. USA* **111,** 18655–18660 (2014).
21. Li, R. *et al.* The sequence and de novo assembly of the giant panda genome. *Nature* **463,** 311–317 (2010).
22. Cahill, J. A. *et al.* Genomic Evidence for Island Population Conversion Resolves Conflicting Theories of Polar Bear Evolution. *PLoS Genet.* **9,** e1003345 (2013).
23. Liu, S. *et al.* Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. *Cell* **157,** 785–794 (2014).
24. Bapteste, E. *et al.* Networks: expanding evolutionary thinking. *Trends Genet.* **29,** 439–441 (2013).
25. Mirarab, S. *et al.* ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30,** (2014).
26. Pease, J. B. & Hahn, M. W. Detection and Polarization of Introgression in a Five-Taxon Phylogeny. *Syst. Biol.* **64,** 651–662 (2015).
27. Huson, D. H. H., Regula, R. & Scornavacca, C. *Phylogenetic Networks*. (Cambridge University Press, 2010).
28. Bidon, T., Schreck, N., Hailer, F., Nilsson, M. & Janke, A. Genome-wide search identifies 1.9 megabases from the polar bear Y chromosome for evolutionary analyses. *Genome Biol. Evol.* **7,** 2010–2022 (2015).
29. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28,** 2239–2252 (2011).
30. Baryshnikov, G. & Zakharov, D. Early pliocene bear *Ursus thibetanus* (Mammalia, carnivora) from Priozernoe locality in the Dniester basin (Molodova republic). *Proc. Zool. Inst. RAS* **317,** 3–10 (2013).
31. Croitor, R. & Brugal, J. P. Ecological and evolutionary dynamics of the carnivore community in Europe during the last 3 million years. *Quat. Int.* **212,** 98–108 (2010).
32. Than, C., Ruths, D. & Nakhleh, L. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics* **9,** 322 (2008).
33. Mailund, T. *et al.* A New Isolation with Migration Model along Complete Genomes Infers Very Different Divergence Processes among Closely Related Great Ape Species. *PLoS Genet* **8,** e1003125 (2012).
34. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475,** 493–496 (2011).
35. García-Rangel, S. Andean bear *Tremarctos ornatus* natural history and conservation. *Mammal Rev.* **42,** 85–119 (2012).
36. Meijaard, E. Craniometric differences among Malayan sun bears (*Ursus malayanus*); evolutionary and taxonomic implications. *Raffles Bull. Zool.* **52,** 665–672 (2004).
37. Edwards, S. V. *et al.* Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. *Mol. Phylogenet. Evol.* **94,** 447–462 (2016).
38. Nakhleh, L. Computational approaches to species phylogeny inference and gene tree reconciliation. *Trends Ecol. Evol.* **28,** 719–728 (2013).
39. Lammers, F., Gallus, S., Janke, A. & Nilsson, M. A. Phylogenetic conflict in bears identified by automated discovery of transposable element insertions in low coverage genomes. arXiv preprint arXiv:123901 (2017).
40. Puckett, E. E., Etter, P. D., Johnson, E. A. & Eggert, L. S. Phylogeographic Analyses of American Black Bears (Ursus americanus) Suggest Four Glacial Refugia and Complex Patterns of Postglacial Admixture. *Mol. Biol. Evol.* **32,** 2338–2350 (2015).
41. Davison, J. *et al.* Late-Quaternary biogeographic scenarios for the brown bear (*Ursus arctos*), a wild mammal model species. *Quat. Sci. Rev.* **30,** 418–430 (2011).
42. Harrison, R. G. & Larson, E. L. Hybridization, Introgression, and the Nature of Species Boundaries. *J. Hered.* **105,** 795–809 (2014).
43. Jarvis, E. D. *et al.* Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346,** 1320–1331 (2014).
44. Nosenko, T. *et al.* Deep metazoan phylogeny: when different genes tell different stories. *Mol. Phylogenet. Evol.* **67,** 223–233 (2013).
45. Hallström, B. M. & Janke, A. Mammalian Evolution May not Be Strictly Bifurcating. *Mol. Biol. Evol.* **27,** 2804–2816 (2010).
46. Suh, A., Smeds, L. & Ellegren, H. The Dynamics of Incomplete Lineage Sorting across the Ancient Adaptive Radiation of Neoavian Birds. *PLoS Biol.* **13,** e1002224 (2015).
47. Wu, C.-I. The genic view of the process of speciation. *J. Evol. Biol.* **14,** 851–865 (2001).
48. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).
49. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **25,** 1754–1760 (2009).
50. Baker, M. De novo genome assembly: what every biologist should know. *Nat. Methods* **9,** 333–337 (2012).
51. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **25,** 2078–2079 (2009).
52. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907v2. (2012).
53. Orlando, L. *et al.* Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499,** 74–78 (2013).
54. Smit, A., Hubley, R. & Green, P. *RepeatMakser* Open-4.0 http:/www.repeatmasker.org (2015).
55. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinforma. Oxf. Engl.* **26,** 841–842 (2010).
56. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet. TIG* **16,** 276–277 (2000).
57. Shimodaira, H. An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* **51,** 492–508 (2002).

58. Stamatakis, A. RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics* **30,** 1312–3 (2014).
59. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* **9,** 772 (2012).
60. Felsenstein, J. PHYLIP (Phylogeny Inference Package) version 3.6. Available from: Author Department of genome sciences, University of Washington. Seattle. (2005).
61. Huson, D. H. & Bryant, D. Application of Phylogenetic Networks in Evolutionary Studies. *Mol. Biol. Evol.* **23,** 254–267 (2006).
62. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15,** 356 (2014).
63. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
64. Yu, Y., Dong, J., Liu, K. J. & Nakhleh, L. Maximum likelihood inference of reticulate evolutionary histories. *Proc. Natl. Acad. Sci.* **111,** 16448–16453 (2014).
65. Onorato, D. P., Hellgren, E. C., van Den Bussche, R. A. & Doan-Crider, D. L. Phylogeographic Patterns within a Metapopulation of Black Bears (*Ursus americanus*) in the American Southwest. *J. Mammal.* **85,** 140–147 (2004).
66. Veeramah, K. R. & Hammer, M. F. The impact of whole-genome sequencing on the reconstruction of human population history. *Nat. Rev. Genet.* **15,** 149–162 (2014).

Acknowledgements

We are grateful to Luay Nakhleh (Rice University) for expert help with Phylo-Net analyses, Yichen Zheng for valuable comments on the manuscript and to Jon Baldur Hlidberg (www.fauna.is), and Aidin Niamir for artwork. Blood samples were kindly provided by Carsten Ludwig (Allwetter Zoo Münster), Tim Schikora (Zoo Schwerin), Christian Wenker (Basel Zoo) and Eva Martinez Nevado (Zoo Madrid). This study was supported by Hesse's funding program LOEWE (Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz) and the Leibniz Society.

Author Contributions

A.J. designed the research and obtained funding. A.J. and T.B. collected the data; V.K. and F.L. conducted the analyses; L.K. provided pedigrees and located samples; A.J., V.K., M.P., M.N., F.L., and T.B. interpreted the results; A.J. and V.K. wrote the paper with the help of all authors.

Additional Information

Accession Codes: The raw reads of the genome sequences have been deposited in the European Nucleotide Archive under the BioProject accession code PRJEB9724.

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing Interests: The authors declare no competing financial interests.

How to cite this article: Kumar, V. *et al.* The evolutionary history of bears is characterized by gene flow across species. *Sci. Rep.* **7**, 46487; doi: 10.1038/srep46487 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.