Genome-wide detection of transposable elements for mammalian phylogenomics

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt beim Fachbereich 15
der Johann Wolfgang Goethe-Universität
in Frankfurt am Main

von

Fritjof Paul Lammers

aus Bochum

Frankfurt, 2019
(D30)

vom Fachbereich 15 der
Johann Wolfgang Goethe-Universität als Dissertation angenommen.

Dekan: Prof. Dr. Sven Klimpel

Gutachter: Prof. Dr. Axel Janke
Prof. Dr. Ingo Ebersberger

Datum der Disputation: 29.08.2019

Contents

Summary

Transposable elements (TEs) are replicating genetic elements that comprise up to 50% of mammalian genomes. A specific class of TEs are retrotransposons that proliferate by transcription into a RNA intermediate, followed by genomic reintegration into another locus (so called “copy & paste” mechanism). Due to the lack of removal mechanisms and very rare parallel insertions, the presence of TE insertions at ortholgous genomic loci in multiple taxa provides a virtually homoplasy-free phylogenetic marker. So far, developing phylogenetically informative markers from TE insertions has been a tedious work of testing hundreds of putative candidate loci in a trial-and-error approach with low success rate. Hence, phylogenetic studies using TE insertions were often limited to a few dozen markers.

Recently, genome sequencing of multiple species using reference-mapping allowed the identification of genome-scale datasets of TE insertions. and made the ad-hoc development of phylogenetic informative markers possible. However, genome-scale TE detection methods have rarely been applied to non-model organisms in which data availability and quality is comparably limited. In this thesis, I developed the TeddyPi pipeline (TE detection and discovery for phylogenetic inference), a software tool that made it possible to obtain reliable genome-scale TE insertion data from low-coverage genomes. This was achieved by integrating the data from multiple TE and structural variation callers as well as applying a stringent filtering pipeline to exclude low-quality insertion calls. Whole-genome sequencing datasets of bears (Ursidae) and baleen whales (Mysticeti) were used to apply TE-based phylogenetic inference and evaluate the method in comparison to sequence-based phylogenomic analyses.

In the bear genomes, TeddyPi identified 150,513 high-quality transposable element (TE) insertions, which allowed me to reconstruct the evolutionary history of bears despite extensive phylogenetic conflict (Lammers et al., 2017). The large number of detected TE insertions made also detailed network analyses possible that visualize the phylogenetic conflict. Experimental polymerase chain reaction (PCR) assays

validated up to 93 % of the computationally identified TE loci and demonstrated the high accuracy of the dataset underlying the phylogenetic analyses.

Second, I present the initial genome sequencing of six baleen whales and a detailed investigation of their evolutionary history using TE insertions and established sequence-based phylogenomic methods. The taxon sampling of baleen whales included iconic species like the blue whale (*Balaneoptera musculus*) or the humpback whale (*Megaptera novaengliae*) (Árnason et al., 2018). A sequence-based reconstruction of the baleen whale species tree solved the long-debated phylogenetic position of the gray whale (*Echrichtius robustus*) within rorquals (Balaneopteridae) for the first time with high statistical support. Furthermore, the genome data made it possible to identify large extent of phylogenetic conflict for divergences during the radiation of rorquals that occurred 7-10 million years ago (Ma).

The phylogenomic analyses of 91,589 TE insertions in the whale genomes confirmed the sequence-based topology (Lammers et al., 2019). The quantification of phylogenetic signals obtained from the TE insertions revealed a high degree of discordance for the divergence of the gray whale and rorquals. Despite the large genome-scale dataset, statistical tests showed only marginal support for a bifurcating divergence of gray whales and the rorqual species. The limited statistical support for a strictly bifurcating tree obtained from genome-scale datasets of thousands of markers demonstrates the importance for including phylogenetic networks for displaying evolutionary divergences.

In conclusion, this thesis shows that identification of TE insertions from whole-genome resequencing provides plentiful and accurate phylogenomic markers. For the application in non-model organisms, I provide a easy-to-use software to integrate multiple datasets from TE and structural variation callers in order to obtain reliable and ascertainment-bias free datasets. Detecting genome-scale datasets of TE insertions in two case studies demonstrates the applicability of this marker system for phylogenetic reconstruction and inferring phylogenetic conflict.

Zusammenfassung

Untersuchungen zur genomweiten Detektion von Retroposoninsertionen im Rahmen phylogenomischer Rekonstruktion der Säugetiere

Hintergrund

Mobile genetische Elemente (MGE), auch genannt Transposable Elemente (TE) sind eigenständig replizierende Regionen im eukaryotischen Genom. Das Genom von Säugetieren besteht bis zu 50 % aus TEs, da die Elemente sich hier über viele Millionen Jahre der Evolution vermehrten. Dabei sind parallele Insertionsereignisse des gleichen Elementtyps sehr unwahrscheinlich und es ist kein molekularer Prozess zur Entfernung integrierter TEs bekannt. Aufgrund dieser Eigenschaften - eindeutige Integration neuer Kopien, sowie dem Fehlen gezielter Entfernungsmechanismen - stellen TE Insertionen phylogenetische Marker dar, die praktisch keine Homoplasie aufweisen. Das seltene Auftreten von Homoplasie bei TEs unterscheidet diese von Punktmutationen oder morphologischen Merkmalen und zeichnet TEs als besonders robuste phylogenetische Marker aus.

Diese Arbeit befasst sich mit den als “Class I” beschriebenen Retrotransposons, da diese diese in den untersuchten biologischen Systemen aktiv sind und sich somit zur Rekonstruktion von Abstammungsverhältnissen eignen. Aktive Retrotransposons in Säugetieren sind das LINE1/L1-Element (Long Interspersed Nuclear Element) und verschiedene Familien von SINEs (Short Interspersed Nuclear Element). L1-Elemente kodieren Proteine, die die Funktionalität zur Retrotransposition bereitstellen. SINEs sind zur Vermehrung von dieser bereitgestellten Funktion abhängig, und werden daher als nicht-autonome Retrotransposososons beschrieben. Der Retrotranspositionsmechanisms verläuft nach einem “Copy & Paste” Schema, nach dem bisherige Kopien des Elements an einen anderen genomischen Locus dupliziert werden.

Die Entwicklung TE-basierter phylogenetischer Marker bedurfte bisher eines zeitaufwendigen molekularbiologisch-experimentellen Aufwands. Dies ist darin begründet, dass es keine universellen genetischen Loci gibt an denen eine große Anzahl

variabler, und damit phylogenetisch informativer, TE Insertionen gibt. Stattdessen müssen über einen Versuch-und-Irrtum Ansatz eine Großteil von Kandidatenloci mittels Polymerase-Kettenreaktion in mehreren Arten getestet werden, um informative Marker zu identifizieren. Neben dem hohen zeitlichen und materiellen Aufwand, war die Anzahl der so identifizierten Markern gering und lag typischerweise bei wenigen Dutzend pro Studie. Somit es erstmals möglich, mit sequenzunabhängigen Markern bestehende stammesgeschichtliche Hypothesen zu beurteilen, in dem die Kongruenz des extrahierten phylogenetischen Signals, mit dem bestehenden Baum überprüft wurde. Aufgrund der vergleichsweise geringen Anzahl informativer Marker, war jedoch die Auflösung komplexer Aufspaltungsereignissen begrenzt nur möglich. Für die detaillierte Aufklärung evolutiver Prozesse, wie tiefer Koaleszenz oder introgressiver Hybridisierung, werden daher mehr informative TE insertionen benötigt.

Mit der größeren Verfügbarkeit von Referenzgenomen, sowie der Möglichkeit zur Resequenzierung weiterer Individuen oder Arten, wurde es möglich, variable TE Loci oder anderer strukturelle genetische Variation direkt zu lokalisieren und zu charakterisieren. Dies geschieht mittels sogenannter “paired-end mapping” Signaturen zusammengehöriger Reads, deren ermittelte Position auf dem Referenzgenom (“Mapping”) eine genetische Variation abbilden. Somit wird es möglich Tausende phylogenetische Marker direkt zu über das gesamte Genom zu identifizieren und damit stammesgeschichtliche Rekonstruktion mittels statistischer *ab initio* Baumsuchverfahren durchzuführen. Die zu erwartende größere Anzahl ermittelter Marker, läßt darüber hinaus auch die Rekonstruktion komplexer Divergenzprozesse zu. Dazu können phylogenetische Netzwerke herangezogen werden. Bis auf eine phylogenetische Studie, die diese Methode innerhalb der Menschenaffen angewendet hat, ist die Anwendung von Resequenzierung-gestützter TE Identifizierung in phylogenetischen Studien bisher nicht untersucht worden.

In dieser Dissertation soll die Identifikation von TE insertion mittels Genom-Resequenzierungverfahren für phylogenetische Studien untersucht werden. Der Schwerpunkt liegt dabei auf Nichtmodellorganismen, da diese in der explorativen Evolutionsforschung von besonderem Interesse sind. Methodisch ist dabei zu beachten, dass Referenzgenomen von Nichtmodellorganismen oft fehlerbehafteter sind und eine weniger ausgearbeitete Charakterisierung aufweisen als Referenzgenomen biomedizinischer Modellorganismen, wie denen des Menschen oder der Hausmaus. Dies ist von Bedeutung, da die Qualität eines Referenzgenoms bedeutenden Einfluss auf die Alignierung von Resequenzierungsdaten hat. Ebenfalls muss die “Tiefe” der Resequenzierungsdaten als mögliche Quelle von technischen Artefakten berücksichtigt werden.

Daher befasst sich diese Arbeit auch mit der Unterscheidung genuiner und fehlerhafter Insertionssignaturen.

Durchgeführte Studien

Die drei Studien, die Bestandteil dieser Dissertation sind, behandeln 1.) die Entwicklung einer Methode zur Identifizierung von phylogenetischen TE Insertionen mittels Hochdurchsatzsequenzierung am Beispiel der Bären (Ursidae), 2.) der Durchführung einer umfangreichen evolutionsgenomischen Studie der Bartenwale (Mysticeti) mittels erprobter Sequenzanalyseverfahren, sowie 3.) die phylogenomische Untersuchung der Bartenwale mithilfe der neu entwickelten TE-Methode. Die zweite und dritte Studie ermöglichen somit einen direkten Vergleich der Ergebnisse von sequenz- und TE-basierten Untersuchungen.

In der ersten Studie entwickle ich zunächst den theoretischen Rahmen für akkurate phylogenetische Rekonstruktion mit TE Insertionen, die genomweit durch Resequenzierung identifiziert werden. Hierzu zeige ich auf, dass zur Vermeidung einer Stichprobenverzerrung ("ascertainment bias"), die Erfassung von TEs auf allen evolutionäre Linien innerhalb der untersuchten Artengruppe, nötig ist. Für den Anwendungsfall, dass das Referenzgenom phylogenetisch innerhalb der untersuchten Artengruppe liegt (oder vermutet wird), bedeutet dies, dass die alleinige Identifizierung von TE Insertionen außerhalb des Referenzgenom (Ref^-) unzureichend ist. Die meisten etablierten Programme zur Identifizerung von TE Insertionen erkennenn jedoch lediglich diese sogenannten Ref^-insertionen. Somit ist die zusätzliche Erkennung von Insertionen, die im Referenzgenom und den resequenzierten Genomen vorkommen, notwendig.

Auf diesem Ergebnis basierend, beschreibe ich die Entwicklung einer bioinformatischen Pipeline, die es ermöglicht TE Insertionen auf sämtlichen evolutionären Linien zu identifizieren. Die Pipeline integriert die Ausgabe verschiedener Programme zu Erkennung von Nichtreferenzinsertionen, sowie weiterer struktureller Varianten. Zur Identifizierung von Referenzinsertionen werden Deletionen in den resequenzierten Genomen genutzt die in Größe und Position, mit im Referenzgenomen annotierten repetitiven Sequenzen übereinstimmen. Ein weiter wichtiger Bestandteil des Programms, ist eine Reihe von Filtern, die Insertionen schlechter Qualität entfernt. Die Filter nutzen hierzu Informationen zu Lücken und Fehlern im Referenzgenom, sowie weitere Parameter aus den resquenzierten Genomen. Bis zu 93% der *in silico* erkannten Insertionsereignisse konnten experimentell mittels Polymerase-Kettenreaktion von 111 Loci in den Bärengenomen bestätigt werden. Insgesamt wurden über 150,000

TE Insertionen erkannt und zur Rekonstruktion der Stammesgeschichte der Bären verwendet.

In der zweiten Studie beschreibe ich die erstmalige Sequenzierung der Genome von sechs Bartenwalen: Blauwal (*Balaenoptera musculus*), Finwal (*B. physalus*), Seiwal (*B. borealis*), Buckelwal (*Megaptera novaengliae*), Grauwal (*Eschrichtius robustus*) und Atlantischer Nordkaper (*Eubalaena glacialis*). Die koaleszenzbasierte Sequenzanalyse von 34,192 genomischen Regionen ermöglichten die Rekonstruktion des ersten phylogenetischen Baums der Bartenwale mit hoher statistischer Unterstützung. Der Stammbaum zeigt, dass der evolutive Ursprung der Grauwale innerhalb Furchenwale (Balaneopteridae) liegt, eine systematische Beurteilung die zuvor unzureichend belegt war. Netzwerk- und Genflussanalysen zeigen eine hohen Grad phylogenetischen Konflikts in den Walgenomen, die teilweise auf Genfluss zwischen Populationen anzestraler Walarten zurückgeführt werden konnte. Weitere populationsgenomische Untersuchungen ergaben, dass einige der rezenten Arten eine vergleichsweise hohe genetische Diversität aufweisen. Diese könnte eine hohe Anpassungsfähigkeit an wechselnde Umweltbedingungen ermöglichen. Weiterhin konnte gezeigt werden, dass anzestrale Walpopulationen im Pleistozäns (2.588 - 0.011 Millionen Jahre vor heute) ein vielfach höhere effektive Populationsgröße aufwiesen, die im Laufe der letzten Million Jahre deutlich abnahm. Für eine detaillierte Bestimmungen heutiger und vergangener populationsgenetische Parameter ist jedoch die Analyse von mehreren Individuen, repräsentativ für die Populationen, notwendig.

Abschließend präsentiere ich die Untersuchung der Bartenwalgenome im Hinblick auf ihre Variabilität von CHR2-Elementen, welche die retrotransponierenden SINEs in Mysticeten darstellen. Es wurde ein Datensatz von 91,589 Loci mit variablen Insertionen erstellt. Phylogenetische Rekonstruktionen bestätigen die in Studie 2 ermittelte Stammesgeschichte und bekräftigten die systematische Einordnung des Grauwals innerhalb der Furchenwale. Netzwerkanalysen des CHR2 Datensatzes bestätigen, dass widersprüchliche phylogenetische Signale vor allem bezüglich der frühen Diversifizierung der Furchenwale auftreten. Statistische Analysen der Verteilung der phylogenetische Signale zeigen dabei, dass die betreffende Divergenz der Grauwale und weiter Furchenwalarten nur mit geringer Untersützung eine Bifurkation darstellt (p=0.0204, KKSC Test). Im Vergleich den heutigen Insertionsraten von CHR2-Elementen, wurde eine höhere anzestrale Insertionsrate ermittelt. Viele dieser anzestralen CHR2 Insertionen sind in den rezenten Genomen noch nicht fixiert und deuten auf hohe effektive Populationgröße des Vorfahren der Furchenwale hin. Somit bestätigt auch diese Untersuchung die vorhergegangen Ergebnisse auf Basis der sequenzbasierten Methoden.

Fazit

In dieser Dissertation konnte ich zeigen, dass die Identifikation variabler TE Insertionen mittels Resequenzierung der Genome von Nichtmodellorganismen eine wertvolle Methode zur Generierung neuer genomweiter phylogenetischer Marker ist. Es konnte dargelegt werden, dass ein hochqualitativer Datensatz durch Integration mehrerer Datenquellen und dem stringenten Filtern wenig reliabler Varianten, erstellt werden kann. Diese Verfahren sind besonders bei “low coverage” Genomen wichtig, um eine möglichst akkurate Darstellung der genomischen Varianten an orthologen Loci abzubilden. Die beiden Fallstudien innerhalb zweier Säugetiergruppen zeigen, dass genomweite Datensätze mit geringen Aufwand erstellt werden können und eine hohe statistische Unterstützung für phylogenetische Verfahren ermöglichen. Gleichzeitig ermöglichen die genomweiten Datensätze, eine detaillierte Bestimmung phylogenetischen Konflikts innerhalb beider Studiensysteme. Die komplementären Untersuchen von Sequenz- und TE-gestützten genomischen Verfahren zeigen erstmalig einen statistisch gestützte Phylogenie für die bisher als kontrovers geltende Stammesgeschichte der Furchenwale und den evolutiven Ursprung des Grauwals. Die Gegenüberstellung der Sequenz- und TE Analysen zeigte eine hohe Übereinstimmung beider Methoden für die phylogenetische Rekonstruktion. Die TE-basierten Analyse ergaben jedoch weniger Signale für interspezifischen Genfluss als die Sequenzanalyse. Mögliche Gründe hierfür sind die geringere Häufigkeit von TE Insertionen gegenüber Punktmutationen sowie des größeren genomischen Impakts einer TE Insertion, die mehrere hundert Basenpaare lang ist. Somit ermöglichen die hier präsentierten Ergebnisse weitergehende Untersuchungen der potentiell differentiell verlaufender evolutiven Prozesse von TE Insertionen und Punktmutationen im Säugetiergenom. Darüber hinaus ermöglichen die hier gewonnen Erkenntnisse weitere Studien im Hinblick auf Adaptationsprozesse die möglicherweise durch Insertionseignisse von TEs ermöglicht wurden, sowie zur weiteren Entwicklung TE-basierte phylogenetische Rekonstruktionsverfahren.

Introduction

Transposable elements are rare genomic changes that are present in all eukaryote genomes. In mammalian genomes TEs represent the major constituent and cover about half of the complete genome sequence (Lander et al., 2001). Their unique mode of proliferation makes TEs ideal phylogenetic markers which are independent from sequence analyses of single nucleotide substitutions (Shedlock et al., 2000; Shedlock et al., 2004). Recent advances in whole genome sequencing (WGS) made it possible to perform genome-wide screens for TE insertions and open the possibility to generate datasets with thousands of phylogenetically informative markers (Medvedev et al., 2009; Durbin et al., 2010). Despite the large potential to identify TE insertions and using them as phylogenetic markers, genome-scale detection of TEs and other structural variations has so far only been applied in a few model mammalian model organisms such as humans, other primates and mice (Stewart et al., 2011; Hormozdiari et al., 2013; Nellåker et al., 2012). Only Hormozdiari et al. (2013) used the identified TE insertions for phylogenetic reconstruction. In part, the limited application of TE detection in evolutionary biology and comparative genomics of non-model organisms is due to the lack of methods for low-coverage sequenced WGS data and a lack of detailed understanding of detection sensitivity and accuracy.

In this thesis, I studied the application of TE and SV calling tools to mammalian genomes that were sequenced at low coverage. The aim was to obtain a genome-scale dataset of reliable phylogenetic markers and compare phylogenomic reconstruction based on TE and sequence analyses. A major discovery of recent phylogenomic studies was that evolutionary histories of many mammalian groups are shaped by extensive periods of genetic exchange by introgressive hybridization and the presence of incomplete lineage sorting (ILS) (Arnold, 2015; Mallet et al., 2016).

Thus, establishing a robust framework to detect TE insertions as genome-scale phylogenetic marker is valuable for understanding complex evolutionary histories and potentially detecting signatures of genetic exchange.

1.1 Mammalian transposable elements

1.1.1 The repetitive genome

Repetitive sequence are the major component of mammalian genomes. In total, over 45% of the genome consist of various types of repeats and only 1.5 % of the 3 billion base pairs are protein coding sequences (Smit, 1999; Lander et al., 2001) (Figure 1.1). The repetitive fraction of the genome is mainly composed of transposable elements (TEs), selfish genetic elements that move and proliferate. Thus, they have considerably contributed to the evolution of mammalian genomes (Deininger et al., 2002; Ivancevic et al., 2016). The activity of TEs during mammalian evolution made a high repeat content and low diversity of the TE composition typical hallmarks of mammalian genomes (Chalopin et al., 2015).

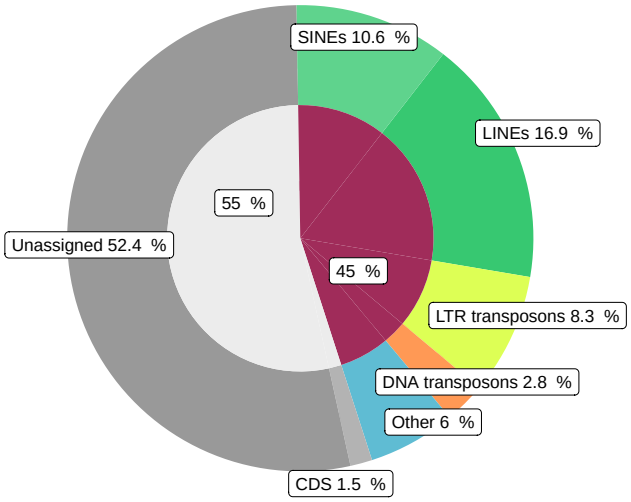


Figure 1.1 Composition of a typical mammalian genome. The inner circle indicates the genomic fraction assigned as repetitive (red) and single-copy (light grey) based on homology methods. The outer circle shows the relative abundances of the repetitive DNA varieties, coding-sequences (CDS) and unidentified genomic sequences. Reproduced from Lander et al. (2001).

The genomic fraction that is neither protein-coding nor categorized as repetitive, likely consists of ancient repetitive sequences, whose copies have accumulated too many mutations to identify their homologs (Lander et al., 2001; de Koning et al., 2011).

Classification of repetitive DNA Transposable elements are classified in a hierarchical scheme that consists of classes, subclasses, families and subfamilies (Jurka et al., 2005). The categorization is based on the propagation mode of the TEs and shared sequence patterns. On the highest level, two classes of TEs are distinguished (Figure 1.2):

Class I are retrotransposons mobilized by transcription into a ribonucleic acid (RNA) intermediate that is reverse transcribed into complementary DNA (cDNA) and integrated elsewhere in the genome. This mode of proliferation can be described as *copy & paste* mechanism and together with the lack of removal mechanisms has lead to the high accumulation of these elements in mammalian genomes. Retrotransposons are further classified by the presence or absence of long terminal repeats (LTRs) flanking the elements (Jurka et al., 2005; Kapitonov et al., 2007). LTR-retrotransposons in most mammals are represented by old copies of endogenous retroviruses (ERVs) and occupy between 4 - 10 % of the genome (Mikkelsen et al., 2007) (Figure 1.1). Non-LTR retrotransposons are the most abundant group of Class I elements and most prominently represented by long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs) that together cover up to 30% of the genome (Lander et al., 2001).

Class II are DNA transposons that mobilize themselves by a Transposase enzyme and generally do not proliferate (*cut & paste*). An exception to this are Helitron transposons that proliferate via a *copy & paste* mechanism (Kapitonov et al., 2007). The hAT DNA transposon was the first TE discovered by Barbara McClintock (1950). In most mammalian genomes, DNA transposons are less abundant than retrotransposons and inactive. Hence, they rarely provide phylogenetic information and are of limited use to reconstruct evolutionary relationships.

For phylogenetic applications in mammals, the retrotransposons SINEs and LINEs are the most commonly used elements due to their almost continuous proliferation during the evolution of mammals. Specifically the autonomous LINE-1 (L1) elements were continuously active during mammalian evolution (Furano, 2004; Sotero-Caio et al., 2017) (see Gallus et al., 2015; Ivancevic et al., 2016 for rare instances of L1 inactivation). L1 elements provided the enzymatic machinery to mobilize non-autonomous SINEs that lack their own coding-capacity and "hitchhike" on the proteins encoded by L1 (Figure 1.3).

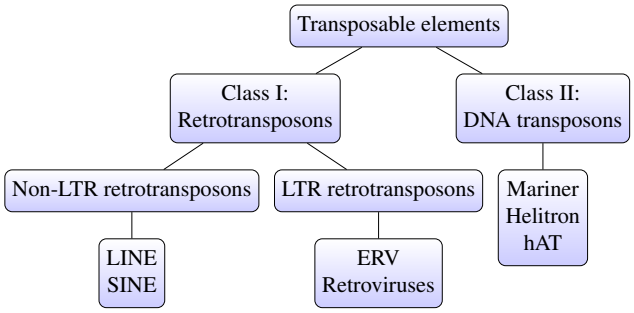


Figure 1.2 Hierarchical classification scheme of the most important TEs in mammalian genomes after Kapitonov et al. (2008).

L1 elements are typically 6 - 7 kilo base pairs (kb) long and have two open reading frames (ORFs) encoding for proteins called ORF1p and ORF2p that mediate retrotransposition (Furano, 2000). ORF2p encodes endonuclease (EN) and reverse transcriptase (RT) domains that are directly responsible for the retrotransposition mechanism (Hattori et al., 1986; Feng et al., 1996). Figure 1.3 illustrates the process called target primed reverse transcription (TPRT). The integration is mediated by the endonuclease domain, that nicks the DNA (target priming) and reverse transcribes the TE-RNA into cDNA (Luan et al., 1993; Cost, 2002). The cDNA serves as template for the synthesis of the second strand by a DNA polymerase. After the completed synthesis, the double-stranded DNA contains the novel integration site with target site duplication (TSD) on both flanks.

The close interrelationship between many SINEs and L1s is based on the similarity of their 3 ' sequence, that serves as a recognition motif for the reverse transcription of (Jurka, 1997; Okada et al., 1997). Furthermore, the sequence similarity of both TEs suggests that SINEs have originated from fusion events of tRNA or 7S-RNA related sequences and a 3' sequence of a L1 element (Deininger et al., 2002). The large variety of SINEs and the difference of their 7S-RNA (e. g. in the primate-specific Alu) or tRNA-related sequence (e. g. in CHR or SINEC elements, see below) indicate that SINEs families emerged independently several times during the evolution of mammals by similar molecular processes (Platt et al., 2018).

In this thesis, the evolutionary history of bears and baleen whales is investigated by identified presence absence patterns of TEs that were active during the species divergences. In bears, the retrotransposing SINEC1_Ame element consists of a Lysine-tRNA related sequence, a $(CT)_n$ microsatellite and a poly-A tail (Vassetzky et al., 2002; Walters-Conte et al., 2011). Its consensus length is 201 base pairs (bp). In the genomes of whales (Cetacea), and their closest terrestrial relatives hippos and ruminants, the CHR-SINE family is active. It is named after their presence in **C**etacea,

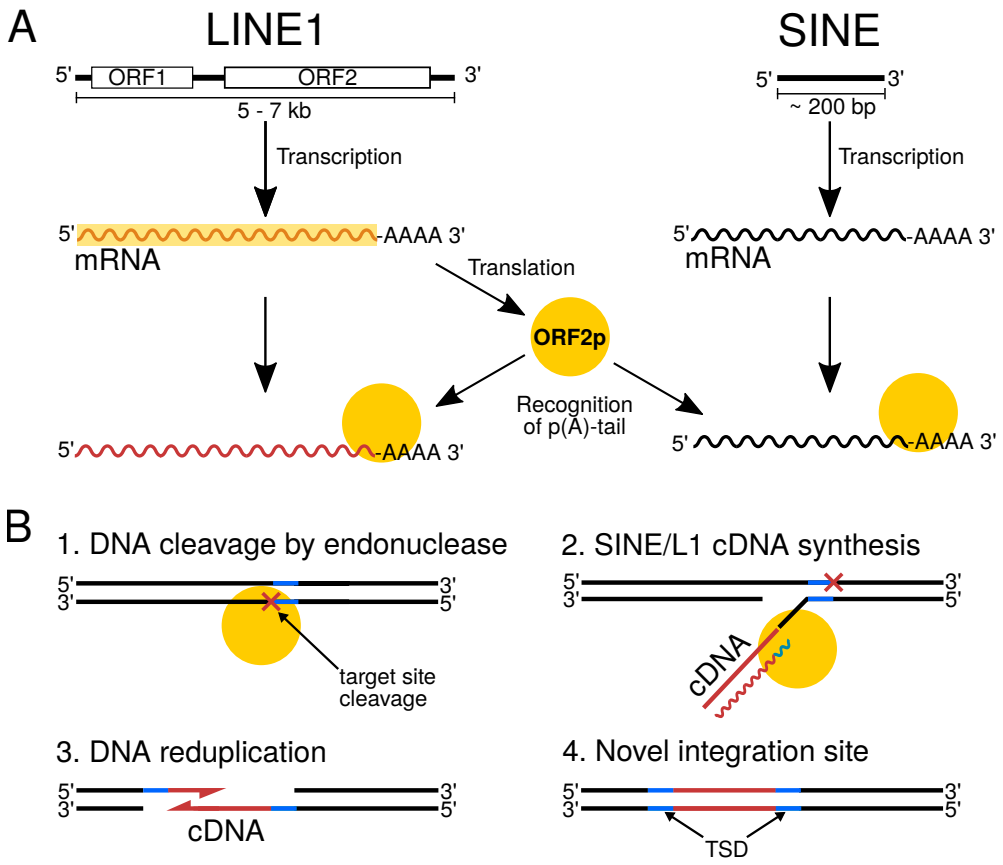


Figure 1.3 Proliferation mechanism of L1 and SINE elements. A) L1 and SINE are transcribed to mRNA. L1 encodes ORF2p (yellow bar/circle)), which provides a reverse transcriptase domain that recognises the poly(A)-tail of L1 and SINE transcripts. B) Target-primed reverse transcription mechansim for L1 and SINE insertions. The endonuclease domain of ORF2p cleaves one DNA strand at the target (1). The SINE or L1 mRNA binds to the exposed single strand and is reverse-transcribed. The second strand is cleaved complementary to the target site. (2). After reverse transcription, the double-strand is reduplicated and the target site duplication (blue) is generated (3,4). Reproduced from Shedlock et al. (2000) and Cordaux et al. (2009).

Hippopotamidae and **R**uminantia. In baleen whales, the 321 bp long CHR2-SINE shows signatures of active retrotransposition (Shimamura et al., 1999). It contains a Glutamine-tRNA related sequence at the 5'end, followed by a 199 bp long tRNA unrelated sequence (Shimamura et al., 1999). Like other mammalian SINEs, SINEC1_Ame and CHR2 elements are mobilized by L1 elements throughout the evolution of their host species. It can be assumed that they substantially contributed to genomic diversity among bears and whales and provide a rich source of phylogenetically informative markers.

1.1.2 Reconstructing phylogenies with TEs

Before the advent of molecular phylogenetics, evolutionary relationships were inferred on the basis of morphological characters that are shared by several species (Hennig, 1965). If such homologous characters are derived from a common original state, they are orthologous and allow to infer common ancestry based on similarity. Shared derived characters that are present in two or more species and originated in their common ancestor are called synapomorphies and are evidence for common descent of the species. This means that highly similar or identical orthologous traits indicate a close evolutionary relationship (Freeman et al., 2007, p. 112).

Molecular phylogenetics

Similar to the comparison of morphological traits, molecular phylogenetics relies on homologous genomic loci that contain synapomorphic characters in order to reconstruct phylogenetic relationships (Williams et al., 2004). Molecular characters can be shared nucleotide sequences, amino acid sequences, TEs insertions or other types of length-polymorphisms. In sequence comparisons, synapmorphies are represented by single nucleotide variants (SNVs), i. e. point mutations in a common ancestor that is subsequently present in at least two taxa (Figure 1.4). In contrast to morphological characters, automated sequencing makes accessing nucleotide sequences easy and cost-effective. Thereby, they provide a large amount of informative characters.

Phylogenetic reconstruction from SNVs and TEs requires specific analytical procedures due to the different molecular mechanisms underlying points mutations and TE integrations. First, for SNVs, the ancestral state is unknown because it can be any of the four nucleotides Adenine (A), Guanosine (G), Cytosine (C) or Thymine (T). Knowing the ancestral state is important to infer the direction of the mutational change and thus to infer phylogenies in a cladistic approach. Outgroup rooting is a method to impute the ancestral state by adding additional taxa that are known to be phylogenetically outside the focal taxon sampling (Maddison et al., 1984). If auxiliary evidence about the phylogenetic position of the outgroup is missing, arbitrary rooting might yield incorrect phylogenies (Figure 1.4).

Second, SNVs are prone to homoplasy, i. e. observing the same character state from independent origin (identity by state). Homoplasy can occur by parallelism, convergence or reversals (back mutations). Parallelisms and convergence are the independent evolution of the same character from the same or different ancestral states, respectively. In both cases, the character states of distantly related taxa appear identical but do not reflect common descent, thus complicating phylogenetic inference.

Reversed sites initially have been synapomorphies, i. e. sites that are identical by descent but subsequently mutated again to the ancestral state. Such recurrent mutations lead to a randomisation of the nucleotide sequence and become problematic for very deep divergences because they can lead to mutational saturation or randomization. Tree reconstruction algorithms account for these processes by incorporating models of sequence evolution that take into account different substitution probabilities and rate heterogeneity (Yang et al., 2012). Still, sequence based phylogenetic inference is left to some approximation and statistical uncertainties by the beforementioned processes. In contrast, TE insertions are virtually free of homoplasy, because parallelisms, convergence and reversals are very rare.

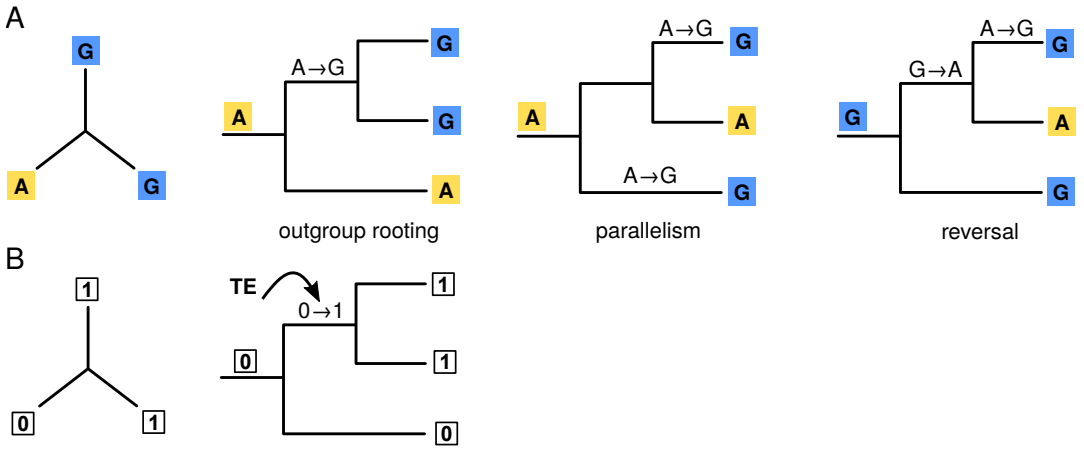


Figure 1.4 Phylogenetic reconstruction of SNVs and TEs. A) From an SNV based unrooted tree (left), different rooted phylogenies might be inferred. The unrooted tree shows that at a given site of three sequences, two different states (A and G) are observed. Depending on whether the ancestral state was A or G, different evolutionary scenarios can explain the observed pattern. The most parsimonious scenario with a single point mutation might not be the correct phylogeny, because also parallelism or reversals might have occurred. B) A genomic locus carrying a TE insertion in two samples is observed. For TEs, the ancestral state (absence of the insertion) is known. Hence, the root for the left tree can intuitively inferred to make a correct phylogenetic reconstruction. Parallelisms and reversals are negligible for TE insertions, because of their rare occurrence. Illustration after Shedlock et al. (2004).

TE-based phylogenetics

As outlined in subsection 1.1.1, retrotranspositionally active TEs shaped mammalian genomes by the continuous integration of new copies. Neutral TE insertions that occur in the germline get inherited to the descendants. Because there is no molecular removal mechanism for TEs, they remain as *genomic fossils*, which indicate common descent when present in several species or populations (Shedlock et al., 2000). Practically, phylogenetic reconstruction from TE insertion requires a presence-absence matrix in

which the status of a synapomorphic TE insertions is coded as present (1), absent (0) or as unknown state (?). The similarity of this data type to morphological characters allows to reconstruct phylogenetic trees using parsimony or distance-based methods. However, due to the limited number of informative markers in early TE based phylogenetic studies, the insertion events were often manually mapped to existing phylogenetic trees to confirm or reject evolutionary hypotheses (i. e. Shimamura et al., 1999; Gallus et al., 2015).

The value of SINE insertions to infer common ancestry of evolutionary lineages was recognized in a study, which presented molecular evidence that the hippopotamus is the closest terrestrial relative to whales (Shimamura et al., 1997). Subsequently, TEs helped illuminating many highly debated divergences in the tree of life as for example baleen whales (Nikaido et al., 2006), toothed whales (Nikaido et al., 2007), primates (Hartig et al., 2013), birds (Suh et al., 2011), marsupials (Nilsson et al., 2010; Dodt et al., 2017) as well as early mammalian lineages (Kriegs et al., 2006; Hallström et al., 2010).

Homoplasious TE insertions are rare because of the relatively slow insertion rate and the huge sequence space, in which TEs can integrate without deleterious effect. In humans, insertion rate are about 0.040 SINEs and 0.006 L1 insertions per generation per genome, respectively (Stewart et al., 2011). Thus, parallel insertions at the exact same locus are stochastically improbable. Empirical analyses of thousands of insertions indicate that parallelisms occur at a rate of 0.0005 per integration event (Ray et al., 2006). Additionally, the lineage-specific evolution of TEs enables researchers to distinguish between independent insertion events by identifying diagnostic mutations in the inserted TE sequence.

Contrary to SNVs that are under constant mutational pressure to change state, there is no known molecular mechanism to excise retrotransposons after integration (Shedlock et al., 2000). Random genomic deletions remove either only fragments of the TE insertion or exceed the integration size and are hence easily identified by sequencing and analysing the particular locus. Recombination of the TSD that flank TE insertions is the only process that can lead to exact removal and occurs with a frequency of 0.005 per insertion event (van de Lagemaat, 2005). TSDs are created during transposition due to the sticky end repair of the nicked integration site (see Figure 1.3).

Furthermore, TE based phylogenetic inference does not require outgroup rooting because the ancestral state of TE insertion is always its absence (Batzer et al., 1994). This make TEs polarized phylogenetic markers and external information to root the phylogenetic tree becomes obsolete (Figure 1.4 B). Finally, observed germline inser-

tions of TEs are likely neutral because deleterious insertions would not be inherited. Thus, phylogenetic reconstruction from TE insertions reflect genetic drift and are not biased by natural selection, which might act on coding and regulatory sequences.

An caveat of previous TE-based phylogenetic studies was the time-consuming work to identify informative markers using PCR and Sanger sequencing. To this end, the typical workflow started with the identification of candidate loci containing TE insertions in one or, at maximum, a few available reference genome sequences. The candidate loci had then to be tested by PCR amplification, gel-electrophoresis screenings and sequencing of orthologous loci in other study species. Only this experimental approach allowed to evaluate the presence or absence of a TE insertions and determine whether the locus is phylogenetically informative. In turn, hundreds to thousands of candidate loci were manually screened to yield not more than several dozens of markers (Lankenau et al., 2009). Shedlock et al. (2004) described this as a "large effort to signal ratio" that impeded large-scale phylogenetic studies based on TEs. Additionally, selecting markers from a single reference genome introduced an ascertainment bias because only loci with a TE insertion present in the reference taxon can be identified (Kuritzin et al., 2016; Dodt et al., 2017). The advent of paired-end sequencing allows to identify TE insertions and other structural variants directly in multiple genomes and offers new ways to rapidly compile genome-scale datasets of informative TE insertions for phylogenetic inference.

1.1.3 Detecting genomic variation

Modern high throughput sequencing technologies allow the increasingly cost-effective generation of whole-genome sequences. In fact, compared to the decade-long billion dollar effort to sequence the first human genome, producing whole-genome datasets representing several closely related species or populations has now become common practice (Prado-Martinez et al., 2013; Jarvis et al., 2014; The 1000 Genomes Project Consortium et al., 2015). In comparative genomics, the typical workflow to detect genomic variation requires the *de novo* assembly of the genome sequence, called a *reference genome*, and additional *resequencing* datasets that represents the genomic information of other individuals or closely related species (Ellegren, 2014).

Currently, the vast majority of available genome assemblies has been generated from short reads produced by Illumina's sequencing-by-synthesis technology. For sequencing, a DNA library is prepared by physically shearing isolated DNA molecules in fragments of 300 - 800 bp and ligating them to adapter sequences. The adapters hybridise to oligonucleotide probes in a flow-cell, where the DNA fragments are

amplified to generate clusters, followed by denaturing of the double stranded DNA. Sequencing is performed by the synthesis of the reverse strand with fluorescent-labeled nucleotides that emit light pulses on incorporation by the DNA polymerase. The light pulses are recorded by a camera and used to reconstruct the DNA sequence. Each fragment is sequenced from both ends, creating so called paired-end (PE) reads, i. e. read pairs consisting of up 150 bp long reads.

Assembling a mammalian reference genome from short reads is challenging because of the large genome size and high repeat content. These parameters complicate the assembly process and leave most genome sequences, at least partially, fragmented and incomplete (Ekblom et al., 2014). Consequently, assembly artefacts require considerations for the mapping of resequencing datasets and the detection of genomic variation (Alkan et al., 2011b; Treangen et al., 2012). Resequencing datasets require a lower depth of coverage (1 - 30X) because the available genomic structure from a closely related reference genome is used. To access the genomic information from resequencing data, the reads are mapped to the reference genome and the variation between reads and the reference genome sequence is examined. This approach allowed to detect millions of single nucleotide polymorphisms (SNPs) (or SNVs for variants between species) among human populations (Durbin et al., 2010) or between great ape (Prado-Martinez et al., 2013) or bear species (Miller et al., 2012; Kumar et al., 2017).

Resequencing with paired-end sequenced DNA fragments made it also possible to detect structural variants and TE insertions from discordantly mapped read pairs (Tuzun et al., 2005; Alkan et al., 2011a). Discordant read pairs deviate from the expected insert size distribution or mapping orientation and originate from structural variation between the resequenced genome and the reference genome. Figure 1.5 A illustrates how structural variants (insertions and deletions) cause paired end mapping (PEM) signatures by decreasing or increasing the mapped distance of a read pair compared to the sequenced fragment size. For example, insertions are indicated by mapped insert sizes that are shorter than expected because the inserted sequence in the sample genome is not present in the reference. The identification of a TE insertion requires the basic insertion signature (Figure 1.5 A), and additionally a database of known TE consensus sequence to determine the type of the integrated sequence (Figure 1.5 B). Reads containing the integrated TE sequence and anchoring flanks, can be partially mapped by clipping non-matching parts (split reads) and help pinpointing the exact genomic coordinate of the insertion site. This approach substantially increased the resolution to discover structural variants < 10,000 bp compared to experimental DNA hybridisation methodologies (Medvedev et al., 2009) and allowed to discover a great

number of structural polymorphisms that were previously unknown in humans (Durbin et al., 2010).

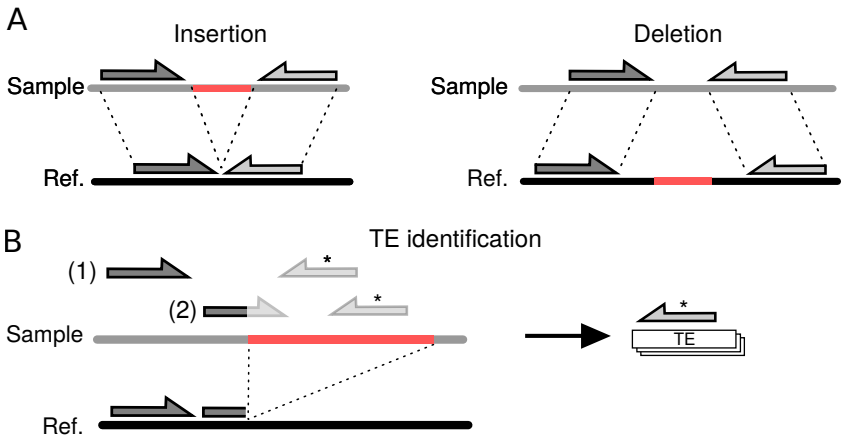


Figure 1.5 Types of structural variants and the corresponding mapping signatures used for their detection. A) Insertions and deletions (red bars) in a resequenced sample and the reference genome (Ref.) cause deviations of the mapped read distance from the original insert size, depicted as dotted lines. Read pairs spanning a insertion show decreased mapped insert size (deletions: increased). B) Identification of TE insertions cause insertion-like mapping signatures and require additional identification of integrated TE-type by comparing the unmapped reads (containig the TE sequence) library of TE consensus sequences (1). Split reads (2) consisting of the integrated TE and flanking sequence are partially mapped and used to determine the exact breakpoint of the TE insertion.

The abovementioned procedure to detect TE insertions from resequencing WGS data increases the number of detected markers compared to PCR based approaches. However, a successful application of TE detection algorithms depends on important parameters such as sequencing coverage, accurate mapping and the quality of the reference genome. The sequencing coverage is the number of bases mapping to each nucleotide position in the reference genome. For TE insertions, the coverage determines how many read pairs support the PEM signature (Ewing, 2015). This is a critical parameter and generally reported by TE calling programs. Sequencing coverage is also the most important parameter for detection sensitivity. For example, the program RetroSeq was reported to have a substantial lower sensitivity below 20X sequencing depth (Keane et al., 2013). Other tools have reported to be more sensitive also at lower sequencing depths (e.g Thung et al., 2014; Gardner et al., 2017), however the higher complexity of PEM signatures compared to SNV calls makes TE detection more error-prone at low sequencing depths.

The high fragmentation, incompleteness and high repeat content of many mammalian genome assemblies can complicate the detection of SV and TE insertions due to inevitable fraction of short reads that are ambiguously mapped (Treangen et al.,

2012). These so called multi-reads are mainly caused by repetitive sequences for which multiple equally good alignments in the genome assembly exist (Li et al., 2010). This problem is aggravated by incompletely assembled genome sequences because the best match identified by the alignment algorithm might be inaccurate if the correct match is missing from the reference genome. Another cause of multi-reads are collapsed repetitive sequences originating from ambiguous cycles in De Bruijn-graph genome assemblies. Thus, genomic repeats increase the probability of misassemblies, assembly gaps and missing sequences, all of which consequently decrease the assembly quality and impede accurate reference-based mapping. Even to the well curated human genome sequence only 70-80 % of resequenced reads can be uniquely mapped because of repetitive and missing sequence (Treangen et al., 2012).

In addition, missing standards and 'Best Practice' guidelines for TE calling are a problem compared to the well defined criteria for successful SNP detection (Van der Auwera et al., 2013; Ewing, 2015). To date, more than 50 implementations of TE and SV detection algorithms exists and were shown to have markedly different performaces (Ewing, 2015; Guan et al., 2016). Combining datasets from multiple programs to complement or cross-validate each other was suggested as possible solution (Wong et al., 2010; Nelson et al., 2017), but none of these workflows has been widely adopted in the field.

1.2 Phylogenomics

The advent of genome sequencing led to the initial expectation that the complete tree of life can get resolved into a succession of bifurcating speciation events if sufficient data becomes available (Philippe et al., 2005). However, many subsequent phylogenomic studies revealed even more phylogenetic incongruencies and showed that these are a natural phenomenon and an important signal of evolutionary processes (Hallström et al., 2010; Bapteste et al., 2013). Processes that lead to gene tree discordance are ILS, selection and introgressive hybridisation. ILS occurs when polymorphisms in the ancestral population of several species remain after the ancestral population diverged into two or more species. Consequently these polymorphisms can indicate a genetic relationship that conflicts with the species history, because non-sister lineages share common alleles. Introgressive hybridisation or gene flow is the result of repeated interspecific hybridisation that led to permanent admixture between two species. The admixture of modern humans and Neanderthals is the most prominent example of ancestral hybridisation revealed by genome sequencing (Green et al., 2010). But also

many other mammals share similar interwoven evolutionary relationships, such as polar and brown bears (Cahill et al., 2013). Hybrids of fin and blue whales have been reported (Árnason et al., 1991; Spilliaert et al., 1991) and raise the question whether these species, too, have hybridised in their evolutionary past. Introgressive hybridization has been considered rare because species are defined by reproductive isolation according to the biological species concept (Mayr, 1963). However, several recent genome-wide studies have shown that the genomes of many species show signatures of interspecific hybridisation and have thereby demonstrated the importance of methods to obtain genome-wide phylogenetic markers (e. g. equids (Jónsson et al., 2014), big cats (Figueiró et al., 2017) or bears (Cahill et al., 2013; Cahill et al., 2016; Kumar et al., 2017)).

1.3 Case studies

1.3.1 Bears (Ursidae)

Bears (Ursidae) are a family that comprise eight extant species in the order Carnivora. The different species are distributed in Eurasia and America and include highly cold-adapted polar bear (*Ursus maritimus*) or generalist species like the brown bear (*Ursus arctos*), which occur in a wide range of habitats (Nowak, 1999). The largest subfamily of bears are the Ursinae, which include six extant species: polar and brown bear, the American black bear (*Ursus americanus*), Asiatic black bear (*Ursus thibetanus*), sloth bear (*Ursus ursinus*) and sun bear (*Ursus malayanus*) (Figure 1.6). The ursine bears evolved in the Pliocene about 3.5-5 million years ago (Mya) (Kutschera et al., 2014; Kumar et al., 2017).

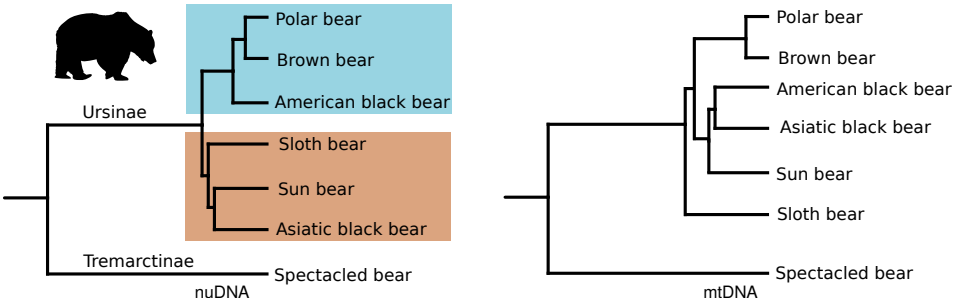


Figure 1.6 Phylogeny of ursine and tremarctine bears estimated from nuclear DNA (nuDNA) and mitochondrial DNA (mtDNA). The subfamily Ursinae evolved in the Pliocene and split into two clades (colored boxes). The conflicting matrilineal phylogeny indicates that these species have undergone several hybridisation events.

All ursine bears live on the northern hemisphere. The spectacled bear (*Tremarctos ornatus*), belonging to the subfamily Tremarctinae, is the only bear species living exclusively on the southern hemisphere and occurs throughout the Andean rainforest in South America. For a long time, the speciation history of ursine bears was unclear because phylogenetic trees reconstructed from mitochondrial DNA (mtDNA) and nuclear markers differed (Krause et al., 2008; Pagès et al., 2008; Hailer et al., 2012; Kutschera et al., 2014). Specifically, the mitochondrial phylogeny placed the polar bear inside brown bears and indicated that the Asiatic and American black bears are sister species. Multi-locus analyses of nuclear introns revealed that polar bears are a distinct lineage that separated from brown bears at least 600 thousand years ago (Hailer et al., 2012). Subsequently, further phylogenomic analyses of more species corroborated that finding and indicated that two distinct clades of ursine bears exist (Figure 1.6) (Kutschera et al., 2014; Kumar et al., 2017). According to these analyses, the American and Asiatic black bear are not sister species. Instead they are grouped separately with other species in the two clades. Specifically, the Asiatic black bear was grouped together with sun and sloth bear, which are the other two bear species that are found in south-east Asia.

The contrasting evolutionary history of the mitochondrial and nuclear genomes, as well as the conflicting phylogenies among nuclear loci must thus be a signal for gene flow and requires further investigation. Hence, adding genome-scale TE insertions as independent phylogenetic markers might give valuable information for the complex evolutionary history of bears.

1.3.2 Baleen whales (Mysticeti)

Baleen whales (Mysticeti) are a systematic group of 15 extant species. They include the largest living animal on earth, the blue whale (*Balaenoptera musculus*) that reaches a body length of up to 30 m and a body mass of 170 metric tons (Nowak, 1999). Baleen whales diverged from toothed whales (Odontoceti) about 30 Mya. In comparison to the toothed whales, baleen whales have lost their teeth during their evolution. The teeth have been replaced by the baleen, arrays of keratine plates that allow the animal to filter-feed large amounts of small prey from the water.

Four extant families belong to the suborder Mysticeti, the right whales (Eubalaenidae), rorquals (Balaenopteridae), and the monotypic families Cetotheriidae (pygmy right whale) and gray whales (Eschrichtiidae) (Figure 1.7). The phylogenetic position of gray whale (*Eschrichtius robustus*) has been debated because phylogenetic studies based on morphological characters showed that they are an distinct evolutionary lin-

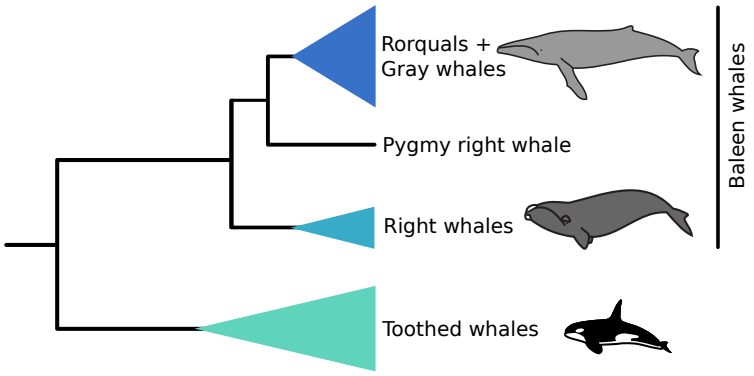


Figure 1.7 Phylogeny of whales (Cetacea). Baleen whales (Mysticeti) are on of the two cetacean suborders and subdivided into several families. Killer whale illustration original by Chris huh (CC BY-SA 3.0).

eage. In contrast, some molecular phylogenetics studies, based on mtDNA and several nuclear markers, indicate a nested phylogenetic position within rorquals (Árnason et al., 2004; McGowen et al., 2009; Steeman et al., 2009; Sasaki et al., 2005). However, the statistical support for this scenario was limited (Steeman et al., 2009). The largest genetic study so far placed the gray whales as sister group to rorquals by using 20 nuclear sequences, 32 TE insertions and several morphological characters in a super-matrix approach (Gatesy et al., 2013). The difficulty in resolving the phylogenetic relationships of baleen whales suggests that evolutionary processes like ILS and gene flow have caused conflicting phylogenetic signals.

Prior to this thesis, the genome sequences of the minke whale (Yim et al., 2014) and the bowhead whale (Keane et al., 2015) have been sequenced in order to yield insight into molecular adaptation to a marine lifestyle and longevity. Thus, additional taxonomically broad WGS resequencing will allow to perform an in-depth genomic analyses of the evolutionary history of baleen whales.

1.4 Thesis objectives

As outlined above, TE insertions can be used as accurate phylogenetic marker because they are virtually homoplasy-free. However, developing phylogenetically markers from TEs is tedious and inefficient using PCR based experimental approaches due to the large number of candidate loci that need to be screened to obtain sufficient numbers of markers. In general, a few dozens of markers were the maximum outcome from this approach.

The aim of this thesis was to establish the genome-wide detection of TE insertions from WGS of non-model organisms in order to develop reliable genome-scale phylogenetic markers and test their applicability in two case studies. It was expected that the genome-wide detection of TE insertions is a valuable method to generate a reliable phylogenetic marker system. that is independent from SNV based tree reconstruction methods. Furthermore, WGS approaches should outperform previous applications of TE insertions in phylogenetics by increasing the number informative markers, minimizing work-intensive experimental testing of candidate loci and eliminating ascertainment bias. In order to test that genome-wide detection of TE insertions can provide an easy-to-generate and ascertainment bias free marker system, I conducted two case studies that investigated the evolutionary history of bears and whales, respectively. Both study system represent phylogenies, that were difficult to resolve in previous studies. Thereby, I could study whether genome scale datasets of TE insertions allow to resolve complex evolutionary histories and detect signatures of introgression and ILS. The parallel phylogenomic analyses based on SNVs enabled me to compare the results of the TE-based phylogenetic reconstruction with established methods.

1. I describe the development of the bioinformatic tool *TeddyPi* (TE detection and discovery of phylogenetic inference) that allows biologists to integrate multiple SV and TE callers and easily design their own stringent filtering pipelines (Lammers et al., 2017, Publication 1). TeddyPi was applied to nine bear genomes in order to create a dataset of TEs insertion that allow to perform phylogenomic analyses of ursine bears and identify phylogenetic conflict caused by introgression or ILS. A companion study provides a detailed sequence-based analyses of the evolutionary history of bears (Kumar et al., 2017) and made it possible compare the phylogenetic signals from both marker systems.

2. I investigated the evolutionary history of baleen whales with SNV and TE based methodologies. To this end, the genomes of six baleen whale species were sequenced for the first time using Illumina paired-end technology. In the first of two studies that used the baleen whales as the study system, I sought to resolve the evolutionary origin of the gray whale among rorquals by coalescent-based multi-locus analyses based on nucleotide sequences. A second aspect of the study was detecting introgression events and investigate the ancestral population size changes using established sequence-based methods (Árnason et al., 2018, Publication 2).

3. Building up on the theoretical work developed for TeddyPi, I generated a genome-scale dataset of TEs for baleen whales with the TE Caller MELT (Gardner et al., 2017). The aim of this study was to independently investigate the phylogeny of baleen whales and identify signals of introgression (Lammers et al., 2019, Publication 3). Thereby, a direct comparison of TE and sequence based phylogenomic analyses will become possible. Additionally, extensive simulations were used to evaluate the performance of MELT for its application to low-coverage genomes.

Discussion

TEs are powerful phylogenomic markers to study complex evolutionary divergence patterns (e.g. Shedlock et al., 2000; Doronina et al., 2017; Dodt et al., 2017). In contrast to sequence-based phylogenetic reconstruction, TE insertions are virtually homoplasy free because they not are prone to convergence, parallelisms, and reversals. In this thesis, I utilized recent advances of high-throughput sequencing technologies that enabled the large-scale detection of TE insertions to establish a framework for phylogenomic reconstruction. The detection of TE insertions from WGS data made it possible to obtain thousands of phylogenetically informative markers instead of a few dozens, which are typical for PCR-based approaches.

In **Publication 1** (Lammers et al., 2017), I developed the theoretical basis for reliable TE detection and accurate unbiased phylogenetic inference. To facilitate the application of TE based phylogenomic studies, I developed the TeddyPi pipeline as a tool that makes it easy to integrate data from multiple variant callers across many samples. As a case study, the application to nine bear genomes successfully extracted 132,039 SINEC1_Ame and 18,420 L1 insertions that enabled a detailed reconstruction of the evolutionary history of ursine bears. Together with Kumar et al. (2017), study allows a direct comparison of TE and sequence based phylogenetic inference. Experimental validation showed that a detection accuracy of over 90% can be achieved even for low-coverage genomes with less than 10X sequencing depth, which makes TE detection a suitable tool for comparative evolutionary genomics.

In **Publication 2** (Árnason et al., 2018), I report the initial sequencing of six new whale genomes as basis for TE and SNV based evolutionary analyses. Similar to other mammalian divergences (e. g. Hallström et al., 2010; Carbone et al., 2014; Kumar et al., 2017), the phylogenomic analyses of baleen whales showed that their evolutionary history represents a phylogenetic network rather than a strictly bifurcating tree. In addition, various signals for gene flow during the radiation of rorquals and gray whale were identified. The distribution of heterozygous sites in the genome sequences

allowed to infer ancestral population sizes, which were much larger during the Pliocene-Pleistocene transition ($\sim$ 2.5 Ma) than contemporary pre-whaling estimates (Roman, 2003; Rocha et al., 2015).

The same genomes were used to study the evolutionary dynamics of TE insertions in rorquals (**Publication 3**, Lammers et al., 2019). This study links the application of genome-scale TE detection in phylogenomics developed in Lammers et al. (2017) to a dataset consisting of 12 baleen whale genomes. Parsimony, distance-based and network analyses of 91,589 identified CHR2 insertion loci reconstructed the phylogenetic history of baleen whales and made a direct comparison of sequence and TE based phylogenetic analyses possible.

In the following discussion, I demonstrate the relevance of genome-scale detection of TE insertions for evolutionary comparative genomics (section 2.1). In section 2.1.2, I describe how high confidence TE calls can be obtained from low-coverage WGS data in order to perform ascertainment bias free phylogenetic inference. This is followed by a discussion of the biological implications of the two case studies (section 2.2) and the significance of these studies in the context of network evolution (section 2.3). Finally, I compare alternative methods for TE detection (section 2.4) and give an outlook to future application of TE in comparative genomics.

2.1 TE insertions as phylogenomic markers

2.1.1 The contribution of TEs to genomic variation

Within both case studies presented in this thesis, tens of thousands of phylogenetically informative TE insertions were identified, each of which can be used as a robust evolutionary marker. In total, more than 240,000 TE insertions were identified in the genomes of bears and baleen whales (Lammers et al., 2017; Lammers et al., 2019). Despite the number of identified TEs is much smaller than the millions of identified SNVs, the potential genetic impact of TEs should not be underestimated considering that a SINE insertion is about 200 - 300 bp long and a L1 element even up to 7 kb. Among human populations, TEs insertions and other SVs cause about 20 Mb of sequence variation, while SNPs account for only $\sim$ 4.1 Mb of genetic difference (The 1000 Genomes Project Consortium et al., 2015). Nevertheless, many studies in comparative genomics only marginally investigate TEs (Goerner-Potvin et al., 2018). Thereby, they neglect a considerable amount of genetic variation that differentiates individuals, populations and species (Sudmant et al., 2015; Prado-Martinez et al., 2013).

The large amount of genetic variation caused by TEs can only occur, because the majority of inherited TE insertions are neutral (Arkhipova, 2018). However, several instances of disease-causing TE insertions were reported in humans (O'Donnell et al., 2010; Cordaux et al., 2009). In other species, TE insertions led to adaptive phenotypes such as shown for the industrial melanism in British peppered moths (van't Hof et al., 2016). The overexpression of TEs in hybrids of lake whitefish (*Coregonus clupeaformis*) suggests that TEs can be involved in speciation processes by causing postzygotic barriers (Dion-Côté et al., 2014). These examples show that studying TE insertions in comparative genomic analyses of natural populations has relevance beyond inferring phylogenies. The large number of TE insertions that can be obtained by genome sequencing provides a great resource for understanding molecular mechanisms of evolution beyond the level of single nucleotides.

2.1.2 The TeddyPi pipeline

Prior to my work, resequencing based TE detection has only been applied within model organisms like mice or great apes, for which well curated reference genome were available (Nellåker et al., 2012; Hormozdiari et al., 2013). Reference genomes of model organisms typically represent the chromosomal structure of the genome, are extensively corrected to avoid misassemblies and contain fewer gaps and missing sequences. For example, the well curated human reference genome represents 95 % of the actual nucleotide content in the genome (Pellicciari et al., 1990; Lander et al., 2001; Peona et al., 2018). In contrast, the genome assemblies from non-model organisms studied here, represent 85% or less of the true genome size because of the underrepresentation of highly repetitive sequences that are often not correctly assembled from short read sequences (Treangen et al., 2012). The polar bear and bowhead whale genome sequences used in this thesis miss about 15% and 22% of the original DNA content (Liu et al., 2014; Keane et al., 2015). Furthermore, a lower continuity (measured by N50, NG50), higher amounts of gaps (N-content), and the presence of misassemblies, such as chimeric joints or collapsed repeats indicate a lower assembly quality (Ekblom et al., 2014; Treangen et al., 2012; Peona et al., 2018).

The limited availability of reference genome sequences might only provide genomes that are phylogenetically nested inside the taxon sampling of the desired phylogenetic study. This can introduce an ascertainment-bias because TE insertions present in the reference are not detected by most TE callers. On this background, Lammers et al. (2017) provides a theoretical framework that defines criteria for ascertainment bias free datasets for phylogenetic inference. The framework is implemented in the

TeddyPi pipeline automatically integrates multiple datasets from WGS based TE and SV detection and filters these data for low-confidence insertions in regions of low assembly and mapping quality.

Overcoming reference bias

In Lammers et al. (2017), I showed that calling TEs from genomes mapped to a reference genome that is phylogenetically nested inside the taxon sampling of a phylogenetic study creates an ascertainment bias. This is because most TE callers only identify TE insertions on evolutionary lineages that exclude the reference genome. These type of insertions are called non-reference (ref^-) insertions (Figure 2.1). From an evolutionary perspective, ref^-TE insertions integrated into genome on lineages that have diverged from the lineage to the reference genome (Figure 2.1 A,B). Hence, it is crucial to identify TE insertions present in (or shared with) the reference in order to obtain an dataset of phylogenetic markers that is free of ascertainment bias. TE insertions shared with the reference genome are called reference insertions (ref^+). Only by detecting both, $ref^+$and ref^-TE insertions, supporting markers for every branch in the phylogeny can be recovered (Figure 2.1). An unbiased representation of phylogenetic markers in the dataset is especially important for the complex evolutionary histories that include reticulations from gene flow events or ILS.

In TeddyPi, the detection of ref^+TE insertions was realized by using deletions calls made by the SV callers Pindel (Ye et al., 2009) and Breakdancer (Chen et al., 2009). Deletions in the resequenced genomes were converted to ref^+TE insertions by comparing their length to known TE sequences and checking for intersection with annotated repeats in the reference genome (Lammers et al., 2017). The identification of 270,689 ref^+compared to 92,196 ref^-insertion calls in the bear genomes demonstrated the enormous contribution of ref^+insertions to the dataset. A similar approach has been applied by Nellåker et al. (2012) to incorporate the complete repertoire of TE insertions across multiple genomes. However the authors did not describe the importance for phylogenetic analyses. It is to note, that SNV-based analyses are not prone to this type of ascertainment-bias because variant calling of single nucleotides can identify all possible character changes, irrespectively of the phylogenetic position of the reference genome species.

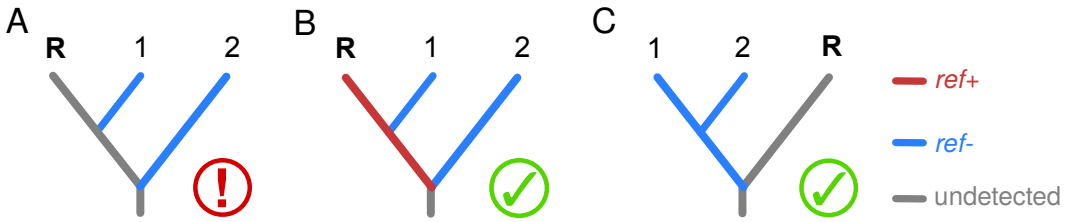


Figure 2.1 Potential introduction of ascertainment-bias in phylogenetic reconstruction by one-sided detection of ref^-TE insertion (blue) if the reference-genome is phylogenetically nested inside the taxon sampling (A). In case of a phylogenetically nested reference-genome, the combination of $ref^-$and ref^+(red) insertions allows to identify insertions on all branches and obtain an ascertainment bias free marker set that can resolve all branches (B). If the reference genome is phylogenetically outside the taxon sampling, ref^-insertions are informative for all branches (C).

Data integration

Obtaining highly reliable datasets of TE insertions from resequenced genomes was possible by combining the results from the TE callers RetroSeq (Keane et al., 2013) and Mobster (Thung et al., 2014), as well as from the SV callers Pindel (Ye et al., 2009) and Breakdancer (Chen et al., 2009). Integrating the results of two ref^-TE callers accounts for the disjunct callsets obtained due to differences in algorithmic and methodological design. For example, Mobster utilizes PEM signatures and split-reads to determine TE insertions (Thung et al., 2014), whereas RetroSeq only used PEM signatures (Keane et al., 2013). Therefore, the incorporation of TE insertions from multiple sources increased sensitivity by about 50%. Similarly, the numbers of detected deletions that were subsequently converted into ref^+insertions differed markedly between the two SV callers (12,865 vs. 296,013) (Lammers et al., 2017). Integrating multiple variant callers has been previously recognized as suitable method to improve variant call accuracy by combining evidence from different variant call algorithms (Wong et al., 2010; Lin et al., 2015). In parallel to the work of this thesis, an automated pipeline for TE detection was developed (Nelson et al., 2017). Compared to TeddyPi, the pipeline by Nelson et al. does not focus on phylogenetic inference and can not incorporate SV calls.

Filtering

TeddyPi uses an implementation of bedtools (Quinlan et al., 2010) to perform the filtering by applying inclusion or exclusion mask equally on all analysed genomes. Thus, TeddyPi enables the user to easily specify filtering steps by editing simple-formatted configuration files. A standard set of filtering-functions is implemented in

the program and new functions can be added via the modular architecture of the tool with little programming knowledge. The following filtering steps are implemented:

Assembly gaps and scaffolds ends In the reference genome sequence, assembly gaps which are represented by 'N's in the sequence (Figure 2.2 A) are an typical assembly artefact. N-regions are introduced during scaffolding with long-range information (such as mate-pairs, linked reads or chromatin-conformation capture) that do not provide the DNA sequence itself, creating gaps filled with an approximate number of 'N's. Hence, the N-regions are problematic, because the exact length of the missed sequence is unknown and short stretches of 'N's might resemble much longer missed sequences and vice versa. In the bear genomes, up to 25% TE calls in direct proximity to N-regions were found and excluded because of their high potential of being false-positives caused by misassemblies around the N-regions.

Nested repeats Another hotspot for assembly artefacts are repetitive sequences that often have lower assembly quality and higher potential of causing ambiguously mapped reads (multi-reads) (Alkan et al., 2011b; Treangen et al., 2012). Repetitive sequences in the genome sequence can cause problems for TE detection algorithms for two reasons. First, repetitive sequences are difficult to assemble *de novo* and might be collapsed or omitted entirely during the assembly process (Treangen et al., 2012; Hoban et al., 2016; Peona et al., 2018). Second, if the same or a similar repeat element integrates into a similar element (nested repeat), multi-read mapping might occur due to the high similarity of the reads of the novel and the old insertion sequence. This can complicate the TE detection process, because no clear PEM signature can be identified.

Mapping artefacts Additionally, mapping artefacts can be identified by deviations from the mean sequencing coverage (Figure 2.2 B). Loci with significant elevated or reduced depth of coverage represent either an copy-number variant site i. e. a biological signature, or accumulations of multi-reads and missing sequences in the sample, respectively. Furthermore, multi-reads can also have invariant sequencing depth and hence might be undetectable by coverage maps.

After applying the filtering steps to all samples and inferring ref^+insertions, the program merges the datasets from all samples to infer the presence and absence status of the TEs and create a unified dataset in BED and NEXUS format. When applied to the nine bear genomes, TeddyPi performed four filtering steps, that were based on

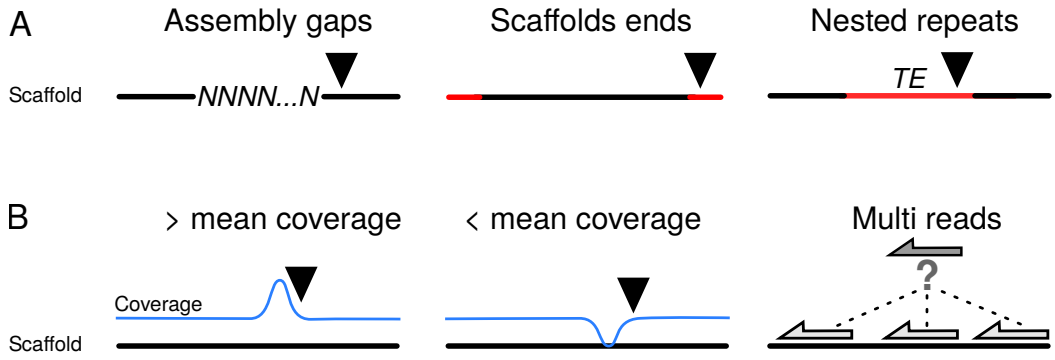


Figure 2.2 Genome sequencing and mapping properties utilized for selecting reliable TE calls. The black arrowhead exemplifies where a TE call is made and excluded. A) Assembly artefacts in genome sequences that decrease TE call quality or produce false-positive variant calls. B) Mapping artefacts of WGS resequencing data.

the TE type, assembly gaps, proximity to other TEs in the reference genome and a coverage map, which excluded regions with sequenced depth < 0.5 and > 2.5X of the mean genome-wide coverage.

Accuracy and sensitivity

The final TeddyPi-generated dataset has been subjected to extensive *in vitro* validation assays (Lammers et al., 2017, Publication 1). These included PCR-amplification and Sanger-sequencing of markers that were selected randomly as well as specifically to test existing phylogenetic hypotheses. The validation experiments showed that resequencing based TE detection achieves high accuracies to more than 90% and generates datasets that can readily be used for phylogenetic inference.

In silico simulations were performed for the TE caller MELT (Gardner et al., 2017) to investigate the effect of sequencing depth on the detection sensitivity and accuracy. To make the simulation experiments comparable to the real data, a error-profile corresponding to the original Illumina data was applied to simulated short reads. The results showed that MELT is very robust to false-positive TE calls even at 5X sequencing depth and also achieves a high sensitivity. The stringent filtering pipeline included in MELT reduced this sensitivity to ∼70%, showing that a high dataset quality is achieved by appyling conservative criteria to exclude low-quality calls (Lammers et al., 2019). However, the joint genotyping of all samples after the initial detection step directly compares the presence absence states of TE insertions at orthologous sites and makes the results of MELT highly reliable for phylogenetic inference. The subsequent processing to identify orthologous insertions in TeddyPi was not necessary.

In summary, reliable TE detection from genomes of non-model organisms is possible if appropriate filtering of low-quality calls and poorly assembled and mapped regions is performed. A pipeline to automate this processing is implemented in TeddyPi (Lammers et al., 2017). Specifically for the case of phylogenetically-nested reference genomes, TeddyPi makes it easy to infer ref^+insertion in order to obtain unbiased datasets. Compared to earlier developed tools, the newly developed TE caller MELT (Gardner et al., 2017) shows a high accuracy at sequencing depths of 5X. This high accuracy comes at a trade off with slightly lower sensitivity. However, for phylogenetic inference a fewer number of correctly identified TE insertions is more important than capturing as many insertions as possible. In addition, joint genotyping of identified insertion loci allows accurate determination of the presence absence status of TE insertions.

2.2 Case studies exemplify the importance of TEs in evolutionary genomics

2.2.1 ILS caused phylogenetic conflict among Asian bear species

The evolutionary history of bears has been enigmatic for long time due to conflicting phylogenies obtained from the mitochondrial and nuclear genomes (Krause et al., 2008; Pagès et al., 2008; Kutschera et al., 2014). Here, the newly developed TeddyPi pipeline generated a dataset of 132,093 SINEC1_Ame and 18,420 L1 insertion loci by the genome-wide screening of seven bear species (Lammers et al., 2017). These datasets made a detailed investigation of the speciation history possible. The phylogenomic analyses unambiguously showed that ursine bear are split into two clades that consists of polar, brown and American black bear as well as of Asiatic black, sun and sloth bear. Therefore, this thesis adds evidence for this evolutionary scenario that has recently been constructed by analyses of nuclear introns and whole-genome sequences (Kutschera et al., 2014; Kumar et al., 2017). Earlier studies used single loci or concatenation of multiple loci to reconstruct phylogenies (e. g. Krause et al., 2008; Pagès et al., 2008). In these studies, the sun bear was placed basal to the other ursine species and the Asiatic black bear was more closely related to the American black bear. The congruence in the new multispecies coalescence and TE-based phylogenies illustrates the necessity for multi-locus analyses in evolutionary studies to efficiently account for gene tree discordance in order to obtain the correct species tree (Edwards, 2009). Because each TE insertion reflects the genealogy of a restricted genomic locus, I

refer to these phylogenetic signals also as "gene trees" in a multispecies-coalescent framework (Maddison, 1997). Therefore, TE insertions are also an independent phylogenetic marker system to perform network analyses.

Phylogenetic networks allow to explore the gene tree discordance from TE insertions, separating the phylogenetic signals from multiple loci (Bandelt et al., 1999). In Lammers et al. (2017), the phylogenetic network showed that the majority of discordance between TE loci indicates ILS, and not gene flow as previously reported (Kutschera et al., 2014; Kumar et al., 2017). Coalescent theory gives an explanation for this observation because in addition to the average waiting time until two lineages coalesce ($2N$ generations), a similar time is required for a TE to integrate due to the slow TE insertion rate per site (Huff et al., 2010). Therefore, TE insertions have on average older coalescences, exhibit deeper genealogies and a higher probability of representing a gene tree that is incongruent to the species tree.

Another explanation for the underrepresentation for the introgression of TE insertions can be the larger genetic impact, which can cause genetic incompatibilities by disrupting coding sequence or affecting gene regulation in a interspecific genetic background. In fact, TEs have been shown to be a major source of regulatory sequence in primate genomes, that co-opted TEs as cis-regulatory elements (Trizzino et al., 2017). Furthermore, TE insertions can indirectly cause structural variation by recombination of homologous and distant TEs, so that introgression of TE loci might be more deleterious than single-nucleotide substitutions (Carvalho et al., 2016). The functional genetic impact of TE insertions in bear genomes requires further research that includes a broader population wide sampling. The detailed analyses of TE insertions might give insights into the speciation processes of bears, which despite their ability to interbreed evolved strikingly different (Kumar et al., 2017; Cahill et al., 2013; Cahill et al., 2015).

2.2.2 Genomic evidence for a rapid radiation of rorquals

In this thesis, whole-genome sequence and TE-based analyses provide a comprehensive view on the evolutionary history of baleen whales (Árnason et al., 2018; Lammers et al., 2019). Two datasets were created that consist of 34,192 genome sequence fragments and 91,859 CHR2 insertions, respectively. From these data, the phylogenetic position of gray whales being inside a paraphyletic rorqual clade was determined for the first time with high statistical support. Previously, this phylogenetic placement has been strongly debated with six out of eleven recent studies finding this divergence unresolved or reporting rorquals to be monophyletic with gray whales being the sister group (Rychel et al., 2004; Árnason et al., 2004; Sasaki et al., 2005; Nikaido et al.,

2006; Demere et al., 2008; Steeman et al., 2009; McGowen et al., 2009; Geisler et al., 2011; Hassanin et al., 2012; Gatesy et al., 2013; Marx et al., 2015). Now, the obtained whole-genome data represents a detailed record of the phylogenetic histories within baleen whale genomes and allowed a comprehensive quantification of the conflict among local genomic genealogies. Figure 2.3 shows the similar strong phylogenetic signals obtained for alternative placements of the gray whale among rorquals.

Based on these conflicting numbers of TE insertions, the statistical support for a bifurcating succession of speciation events inside rorquals was very limited on basis of the TE dataset (p=0.0204, KKSC-test) (Lammers et al., 2019). Similar, the multi-species coalescent tree generated by ASTRAL (Mirarab et al., 2014) from 34,192 genome fragment trees reported equal frequencies of alternating quartet trees for the divergences including the gray whale (Árnason et al., 2018). The sequence and gene tree based methods D-statistics (Durand et al., 2011), D-Foil (Pease et al., 2015) and PhyloNet (Yu et al., 2014) detected various gene flow signals, that must have been taken place between ancestral rorquals species.

Taken together, the highly heterogeneous landscape of gene trees indicate incomplete lineage sorting or ancestral genetic exchange during the speciation processes of rorquals. Speciation with genetic exchange has been reported for many species groups (Arnold, 2015). The connectivity of the marine environment makes migration between populations for the highly mobile baleen whales easy. In fact, gray whales have been reported to migrate between the Pacific and Atlantic ocean, demonstrating the ability for long migrations (Alter et al., 2015). Hence, it can be assumed that geographic boundaries are virtually non-existent for highly migratory species like baleen whales. In addition, it is known that baleen whales share an identical karyotype (2n=44 chromosomes), which facilitates hybridisation (Árnason, 1974) and in fact, blue and fin whale have been reported to hybridise (Spilliaert et al., 1991).

In conclusion, studying the evolutionary history of whales from nucleotide sequence changes and TE insertions indicate that the radiation of rorquals and gray whales is not a bifurcating divergence. The independently obtained datasets of phylogenetic markers indicate equally strong signals for three alternating signals. This can be explained by ILS or speciation with genetic exchange, likely mediated by large ancestral populations with high connectivity.

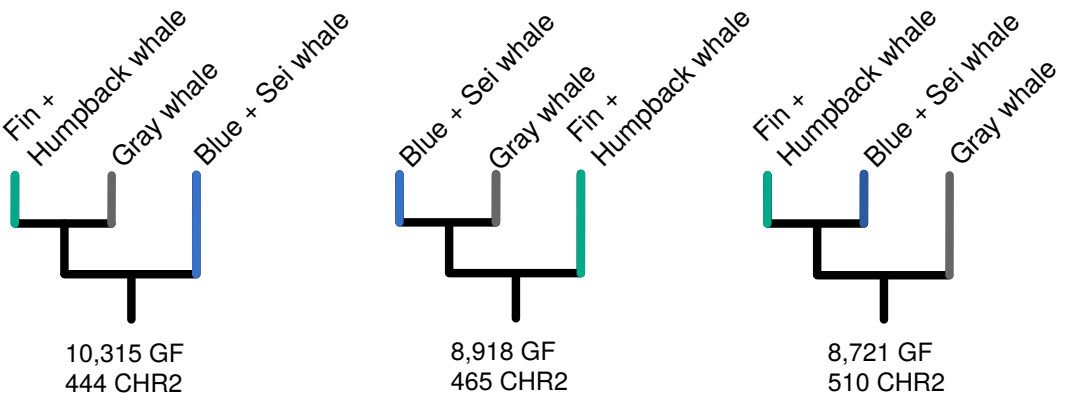


Figure 2.3 Phylogeny of rorquals and gray whales based on a sequence based study from 34,192 genome fragments and a TE based study using 91,859 CHR2 insertions. The trees show the different placement of the gray whale among the other rorquals with the number of supporting genome fragments (GF) and CHR2 insertions for the respective phylogeny.

2.3 Network evolution - phylogenetic discordance is a signal

Studying the evolutionary history of bears and whales, this thesis adds to the accumulating genomic evidence showing that reticulate evolutionary histories are more common than previously thought (Nosil, 2008). Complex evolutionary histories, characterized by a high extent of gene tree discordance or identified gene flow signals were, among others, also found in humans (Slatkin et al., 2016), equids (Jónsson et al., 2014), and bats (Tsagkogeorga et al., 2013). The recently increasing number of such reports demonstrates the value of genome sequencing and multi-locus analyses to yield unprecedented insights into mammalian speciation processes. Also, these findings have consequences for our understanding of the evolutionary tree of life. If not all divergences are bifurcating (Hallström et al., 2010), phylogenetic networks might be a more correct method to display the evolutionary history of species encoded in their genomes (Bapteste et al., 2013). Multifurcation might need to be considered as valid evolutionary scenarios in species trees if near-instantaneous radiation do not allow to reconstruct a clearly bifurcating tree (Árnason et al., 2018; Lammers et al., 2019). Even though multifurcations have been seen as artefacts from insufficient data, new genomic data and geological scenarios such as the concurrent emergence of marine lakes by geological erosion (Maas et al., 2018) provides plausible hypotheses for rapid radiations that are not accurately represented by entirely bifurcating phylogenies.

Acknowledging that local genomic genealogies can be different also has implications for our understanding of the evolution of organismic traits, that is based on

accurate speciation histories (Hahn et al., 2016). For example, gene flow can explain the paraphyletic occurrence of traits if the genes providing functionality of that traits are introgressed. In contrast, if no introgression took place, the trait must have been evolved multiple times, which is considered less parsimomious. Therefore, the incorporation of reticulate evolutionary history allow to infer more parsimonious, and thus more probable, evolutionary scenarios, because gene

Finally, an area of active research is the question whether introgressed alleles can be adaptive. Hybridisation allows to introduce new alleles into a population, thus increasing the genetic variation of the population allowing for higher probability of evolutionary innovations (Abbott et al., 2013). Among mammals, the North American snowshoe hare provides a striking case of introgression leading to an adaptation to changing environmental conditions (Jones et al., 2018) . For the hares, which grow a camouflage white coat during winter, it was found that in areas of lower snow cover also winter-brown individuals are observed. By genome-wide sequencing, population genetic and phylogenetic analyses, Jones et al. (2018) identified the introgression of an *Agouti* allele from black-tailed jackrabbits into snowshoe hares facilitating this local adaptation. Hence, investigating introgression events allow to gain further insights into evolutionary processes and understanding adaptation to changing environmental conditions.

The datasets of genome-wide phylogenetic markers as well as the reconstructed species trees of bears and whale presented in this thesis provide a rich resource to expand future research towards an in-depth understanding of the evolution of these species and their remarkable adaptations to various terrestrial and marine environments.

2.4 TE detection methods in evolutionary biology

Since TEs were introduced as phylogenetic markers more than 20 years ago (Shimamura et al., 1997), available technologies shaped the methodology to detect informative loci (Figure 2.4). In the beginnings of TE based phylogenetic studies, DNA-DNA hybridization using known TE fragments was used in order to identify positive hybridization clones in one species (Figure 2.4 A). Then, PCR primers in flanking sequences of the TE were designed to perform multi-species PCR and sequencing assays that obtained presence or absence patterns of the TE insertion (Shimamura et al., 1997; Nikaido et al., 2006).

With the increasing availability of full genome sequences, the computational screening of TE loci allowed the efficient identification of many candidate loci. These

were tested on a broader taxon sampling with PCR assays (Figure 2.4 B) (Meyer et al., 2012; Gallus et al., 2015; Churakov et al., 2009). Similar, to the one-sided detection of ref^- detection using WGS data, using a single reference genome can introduce an ascertainment bias because it only finds ref^+ insertions (Kuritzin et al., 2016).

Recently, computational multi-genome comparisons became possible if genome assembly for multiple evolutionary lineages in the studied clade were available. For instance, this was the case in a study of the early radiation of birds (Suh et al., 2015), where a complete *in silico* screen of 48 genomes detected thousands of informative TE markers (Figure 2.4 C). This approach is not prone to ascertainment bias because the presence absence status of TE insertions can be inferred directly from the sequences. The caveat of this approach is the lack of reliable algorithms to mine the sequence alignments for TE insertions. This makes it still necessary to manually inspect thousands of candidate loci (Suh et al., 2015).

ME-Scan (Witherspoon et al., 2013) is a targeted sequencing approach that selectively amplifies junctions of a specific TE sequence (Figure 2.4 D). In a phylogenetic study of Myotis bats, *ME-Scan* successfully identified 85,000 informative TE insertions (Platt et al., 2015). Hence, this method successfully identifies large amounts of insertions but only does so in the studied genomes, which introduces bias (Platt et al., 2015). The *STRONG* method uses genome-skimming i. e. ultra low-coverage short-read sequencing ($\sim$ 1X depth of coverage) to obtain a representative fraction of reads that contain TE-junctions (Figure 2.4 E). The flanking regions at the TE-junctions are then used to design primers in order to perform multi-species PCR assays as described above (Kuramoto et al., 2015). Screening multiple genomes with *STRONG* is inexpensive due to low-coverage required and reduces potential ascertainment bias that would be introduced when screening a single species. However, manual PCR-assays are still required to identify informative loci and limit the total number of markers that can be obtained.

In summary, in order to obtain unbiased phylogenetic markers, whole-genome resequencing based TE detection as used by TeddyPi (Figure 2.4 F) is the most practical approach. Compared to the other methods, WGS-based detection enables to genotype the complete taxon sampling and yield high numbers of markers. So far, short reads are the most cost-effective method to identify genetic variation. With decreasing sequencing prices for long reads, inferring variation from these will allow to substantially improve detection of TE insertions because the reads span the complete repeat element (Cretu Stancu et al., 2017). Furthermore, long reads and linked reads such as *10X Chromium* makes the assembly of complete genome sequence increasingly simple and provides better assembly quality (Ma et al., 2018). Hence, the development

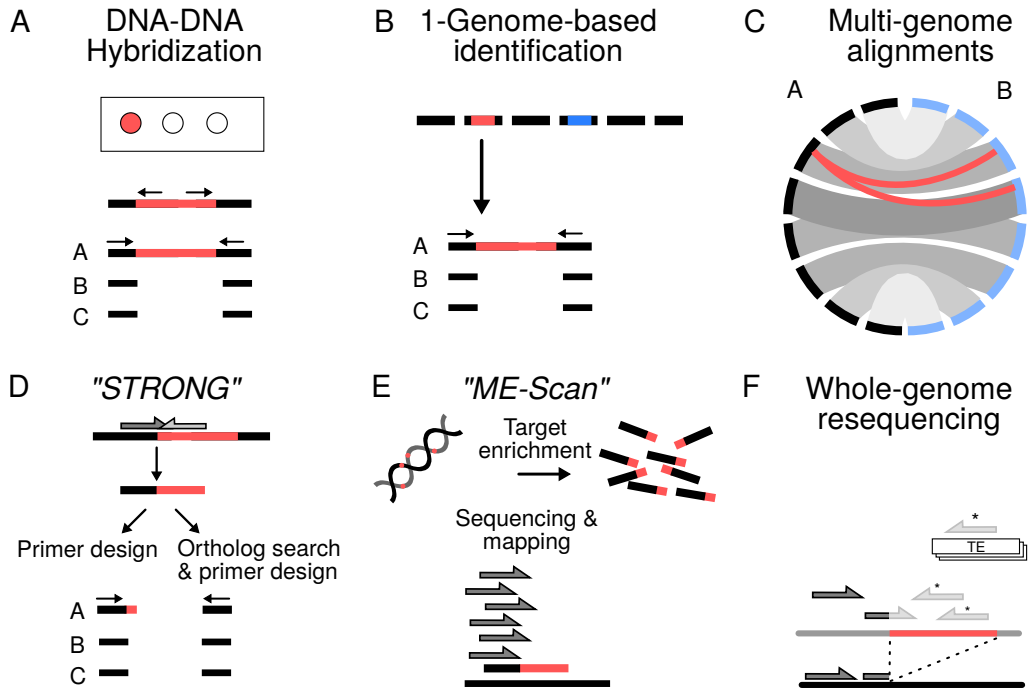


Figure 2.4 Comparison of TE detection methods. A) Loci containing TE sequences are identified in DNA libraries by DNA-DNA hybridisation. Positive hybridisation clones are the basis of primer design and multi-species PCR assays. B) Available genome sequences are screened for TE insertions and primers are designed in conserved flanking regions for multi-species PCR assays as in A. C) Computational analyses of multi-genome alignments to directly identify phylogenetically informative insertions sites across multiple species. D) Genome skimming sequencing to identify TE-flanking regions of over-represented TEs in the read set. Overlapping reads are used for primer design and PCR assays. E) TE-enriched DNA libraries are sequenced across multiple species to identify presence absence of TEs by read mapping to a reference genome. F) Whole-genome sequencing based TE detection by paired-end mapping signatures as utilised by the TeddyPi pipeline.

of efficient algorithms for multi-genome alignment and the detection informative TE insertions have the potential to facilitate TE-based phylogenomic based on genome assemblies.

2.5 A note on phylogenetic algorithms for TEs

Phylogenetic reconstruction from TE insertion is based on a presence absence matrix that indicates the status of the TE insertion at each genotyped locus. Only the advent of genome-wide scans for TE insertions allowed to obtain presence absence matrices large enough to perform *ab initio* phylogenetic inference with high statistical support.

The best fitting method to infer phylogenetic trees from TE inserion datasets, is the Dollo-Parsimony criterion. Under Dollo-Parsimony, the tree with the least number of character gains from the ancestral state (absence) is considered the most parsimonious. In contrast to the standard parsimony method, the Dollo criterion allows every derived character to evolve only once and to rarely get lost (Farris, 1977). Distance-based methods such as Neighbour Joining (Saitou et al., 1987) are another valid approach to infer relationships from the presence absence information of TEs. Maximum likelihood methods that use markov chains as substitution models are not applicable for TE insertion data, because the insertion process is not "memoryless" and hence not a Markov property.

Median or neighbour-net networks visualise the phylogenetic conflict present in a binary presence-absence matrix and were widely applied in sequence and TE-based phylogenomics e.g. (Suh et al., 2015; Suh et al., 2017; Hallström et al., 2010; Kumar et al., 2017). Phylogenetic networks are valuable tools to explore the data but they do not provide any evolutionary analyses to identify possible sources of the phylogenetic conflict. In contrast to the many tools for detecting gene flow from whole-genome sequences, for TE datasets, only the KKSC-test allows to detect gene flow by comparing insertion counts for alternating quartet trees at single branches (Kuritzin et al., 2016).

In the future, the multi-species coalescent (MSC) model might provide an avenue for analysing complex phylogenies from TE and detecting gene flow signals. MSC methods aim to find the species tree that incorporate the coalescent events from multiple loci across the genomes. Thereby, the MSC takes into account ILS and yields a more exact reconstruction of the speciation history than concatenation methods that largely ignore phylogenetic conflict in the dataset. TE insertions could be suitable markers for coalescent based approaches, because they occur interspersed throughout the genome and exhibit locally restricted genealogies. Therefore, a possible method is to consider phylogenetically informative TE insertion as a bipartition, i. e. a phylogenetic tree that separates two groups of taxa. In fact, species tree reconstruction from bipartitions under the MSC has been demonstrated (Allman et al., 2013; Allman et al., 2017). If this concept can be successfully transferred to TE datasets, even more accurate phylogenomic reconstruction and the inference of gene flow events using rooted phylogenetic networks might be possible (Solís-Lemus et al., 2017).

2.6 Conclusion

Establishing a framework for the genome-scale detection of TE insertions from low-coverage WGS datasets, and my applying it in two case studies, I provide an independent phylogenomic marker system to reconstruct accurate species trees.

I developed the TeddyPi pipeline, an easy-to-use application that makes it possible to generate large datasets of accurate presence absence information of TE insertions in multiple species (Lammers et al., 2017). Applying TeddyPi to the genomes of bears, I generated the first datasets of phylogenetically informative TE insertions in these species and reconstruct evolutionary history independently from sequence-based analyses.

Phylogenomic reconstruction of baleen whales with TE-based methods as well as established sequence analyses further showed the value of TE insertions to infer complex evolutionary histories (Árnason et al., 2018; Lammers et al., 2019). Detecting TE insertions in bears and baleen whales I showed that these structural variants contribute considerably to genetic variation that should not be neglected in comparative genomic analyses (Lammers et al., 2017; Lammers et al., 2019). The detection of TE insertions and the TeddyPi pipeline is applicable to all species, for which accurate mapping of short-reads to a reference genome is feasible. Therefore, this thesis makes an important contribution to the methodological repertoire of evolutionary genomics across the whole tree of life. Moreover, the initial sequencing of six baleen whale genomes, provides an important resource for further research in evolutionary biology and conservation genomics.

References

Abbott, R. et al. (2013). "Hybridization and Speciation". *Journal of Evolutionary Biology* 26.2, pp. 229–246.

Alkan, Can, Bradley P Coe, and Evan E Eichler (2011a). "Genome Structural Variation Discovery and Genotyping." *Nature Reviews Genetics* 12.5, pp. 363–76.

Alkan, Can, Saba Sajjadian, and Evan E Eichler (2011b). "Limitations of Next-Generation Genome Sequence Assembly". *Nature Methods* 8.1, pp. 61–65.

Allman, Elizabeth S., James H. Degnan, and John A. Rhodes (2013). "Determining Species Tree Topologies from Clade Probabilities under the Coalescent". *Molecular Phylogenetics and Evolution* 66.3, pp. 628–644.

Allman, Elizabeth S., James H. Degnan, and John A. Rhodes (2017). "Split Probabilities and Species Tree Inference under the Multispecies Coalescent Model". *arXiv [q-bio]* 1704.04268.

Alter, S. Elizabeth et al. (2015). "Climate Impacts on Transocean Dispersal and Habitat in Gray Whales from the Pleistocene to 2100". *Molecular Ecology* 24.7, pp. 1510–1522.

Arkhipova, Irina R (2018). "Neutral Theory, Transposable Elements, and Eukaryotic Genome Evolution". *Molecular Biology and Evolution* 35.6. Ed. by Sudhir Kumar, pp. 1332–1337.

Árnason, Úlfur (1974). "Comparative Chromosome Studies in Cetacea." *Hereditas* 77.1, pp. 1–36.

Árnason, Úlfur, Anette Gullberg, and Axel Janke (2004). "Mitogenomic Analyses Provide New Insights into Cetacean Origin and Evolution". *Gene* 333, pp. 27–34.

Árnason, Úlfur, Fritjof Lammers, Vikas Kumar, Maria A. Nilsson, and Axel Janke (2018). "Whole-Genome Sequencing of the Blue Whale and Other Rorquals Finds Signatures for Introgressive Gene Flow". *Science Advances* 4.4, eaap9873.

Árnason, Úlfur, Rémi Spilliaert, Ástrídur Pálsdóttir, and Alfred Árnason (1991). "Molecular Identification of Hybrids between the Two Largest Whale Species, the Blue Whale (Balaenoptera Musculus) and the Fin Whale (B. Physalus)". *Hereditas* 115.2, pp. 183–189.

Arnold, Michael L. (2015). *Divergence with Genetic Exchange*. New York: Oxford University Press. 272 pp.

Bandelt, Hans Jürgen, Peter Forster, and A. Rohl (1999). "Median-Joining Networks for Inferring Intraspecific Phylogenies". *Molecular Biology and Evolution* 16.1, pp. 37–48.

Bapteste, Eric et al. (2013). "Networks: Expanding Evolutionary Thinking". *Trends in Genetics* 29.8, pp. 439–441.

Batzer, M. A. et al. (1994). "African Origin of Human-Specific Polymorphic Alu Insertions." *Proceedings of the National Academy of Sciences* 91.25, pp. 12288–12292.

Cahill, James A. et al. (2013). "Genomic Evidence for Island Population Conversion Resolves Conflicting Theories of Polar Bear Evolution". *PLoS Genetics* 9.3, e1003345.

Cahill, James A. et al. (2015). "Genomic Evidence of Geographically Widespread Effect of Gene Flow from Polar Bears into Brown Bears". *Molecular Ecology* 24.6, pp. 1205–1217.

Cahill, James A, André E.R. Soares, Richard E Green, and Beth Shapiro (2016). "Inferring Species Divergence Times Using Pairwise Sequential Markovian Coalescent Modelling and Low-Coverage Genomic Data". *Philosophical Transactions of the Royal Society B: Biological Sciences* 371.1699.

Carbone, Lucia et al. (2014). "Gibbon Genome and the Fast Karyotype Evolution of Small Apes". *Nature* 513.7517, pp. 195–201.

Carvalho, Claudia M. B. and James R. Lupski (2016). "Mechanisms Underlying Structural Variant Formation in Genomic Disorders". *Nature Reviews Genetics* 17.4, pp. 224–238.

Chalopin, Domitille, Magali Naville, Floriane Plard, Delphine Galiana, and Jean-Nicolas Volff (2015). "Comparative Analysis of Transposable Elements Highlights Mobilome Diversity and Evolution in Vertebrates". *Genome Biology and Evolution* 7.2, pp. 567–580.

Chen, Ken et al. (2009). "BreakDancer: An Algorithm for High-Resolution Mapping of Genomic Structural Variation". *Nature Methods* 6.9, pp. 677–681.

Churakov, Gennady et al. (2009). "Mosaic Retroposon Insertion Patterns in Placental Mammals". *Genome Research* 19.5, pp. 868–875.

Cordaux, Richard and Mark A Batzer (2009). "The Impact of Retrotransposons on Human Genome Evolution." *Nature Reviews Genetics* 10.10, pp. 691–703.

Cost, G. J. (2002). "Human L1 Element Target-Primed Reverse Transcription in Vitro". *The EMBO Journal* 21.21, pp. 5899–5910.

Cretu Stancu, Mircea et al. (2017). "Mapping and Phasing of Structural Variation in Patient Genomes Using Nanopore Sequencing". *Nature Communications* 8.1.

De Koning, A. P Jason, Wanjun Gu, Todd A. Castoe, Mark A. Batzer, and David D. Pollock (2011). "Repetitive Elements May Comprise Over Two-Thirds of the Human Genome". *PLoS Genetics* 7.12, e1002384.

Deininger, P. L. and Mark A Batzer (2002). "Mammalian Retroelements". *Genome Research* 12.10, pp. 1455–1465.

Demere, T. A., M. R. McGowen, A. Berta, and J. Gatesy (2008). "Morphological and Molecular Evidence for a Stepwise Evolutionary Transition from Teeth to Baleen in Mysticete Whales". *Systematic Biology* 57.1, pp. 15–37.

Dion-Côté, Anne-Marie, Sébastien Renaut, Eric Normandeau, and Louis Bernatchez (2014). "RNA-Seq Reveals Transcriptomic Shock Involving Transposable Ele-

ments Reactivation in Hybrids of Young Lake Whitefish Species". *Molecular Biology and Evolution* 31.5, pp. 1188–1199.

Dodt, William G., Susanne Gallus, Matthew J. Phillips, and Maria A. Nilsson (2017). "Resolving Kangaroo Phylogeny and Overcoming Retrotransposon Ascertainment Bias". *Scientific Reports* 7.1, p. 16811.

Doronina, Liliya et al. (2017). "Speciation Network in Laurasiatheria: Retrophylogenomic Signals". *Genome Research*, gr.210948.116.

Durand, Eric Y., Nick Patterson, David Reich, and Montgomery Slatkin (2011). "Testing for Ancient Admixture between Closely Related Populations". *Molecular Biology and Evolution* 28.8, pp. 2239–2252.

Durbin, Richard M. et al. (2010). "A Map of Human Genome Variation from Population-Scale Sequencing". *Nature* 467.7319, pp. 1061–1073.

Edwards, Scott V. (2009). "Is a New and General Theory of Molecular Systematics Emerging?" *Evolution* 63.1, pp. 1–19.

Ekblom, Robert and Jochen B. W. Wolf (2014). "A Field Guide to Whole-Genome Sequencing, Assembly and Annotation". *Evolutionary Applications* 7.9, pp. 1026–1042.

Ellegren, Hans (2014). "Genome Sequencing and Population Genomics in Non-Model Organisms". *Trends in Ecology & Evolution* 29.1, pp. 51–63.

Ewing, Adam D. (2015). "Transposable Element Detection from Whole Genome Sequence Data". *Mobile DNA* 6.1.

Farris, James S. (1977). "Phylogenetic Analysis Under Dollo's Law". *Systematic Zoology* 26.1, pp. 77–88.

Feng, Q., J. V. Moran, H. H. Kazazian, and J. D. Boeke (1996). "Human L1 Retrotransposon Encodes a Conserved Endonuclease Required for Retrotransposition". *Cell* 87.5, pp. 905–916.

Figueiró, Henrique V. et al. (2017). "Genome-Wide Signatures of Complex Introgression and Adaptive Evolution in the Big Cats". *Science Advances* 3.7, e1700299.

Freeman, Scott and Jon C. Herron (2007). *Evolutionary Analysis*. 4. ed., Pearson internat. ed. OCLC: 255382369. Upper Saddle River, NJ: Pearson/Prentice Hall. 834 pp.

Furano, A (2004). "L1 (LINE-1) Retrotransposon Diversity Differs Dramatically between Mammals and Fish". *Trends in Genetics* 20.1, pp. 9–14.

Furano, Anthony V (2000). "The Biological Properties and Evolutionary Dynamics of Mammalian LINE-1 Retrotransposons". *Progress in Nucleic Acid Research and Molecular Biology* 64, p. 40.

Gallus, Susanne, Axel Janke, Vikas Kumar, and Maria A. Nilsson (2015). "Disentangling the Relationship of the Australian Marsupial Orders Using Retrotransposon and Evolutionary Network Analyses". *Genome Biology and Evolution* 7.4, pp. 985–992.

Gardner, Eugene J. et al. (2017). "The Mobile Element Locator Tool (MELT): Population-Scale Mobile Element Discovery and Biology". *Genome Research* 27.11, pp. 1916–1929.

Gatesy, John et al. (2013). "A Phylogenetic Blueprint for a Modern Whale". *Molecular Phylogenetics and Evolution* 66.2, pp. 479–506.

Geisler, Jonathan H., Michael R. McGowen, Guang Yang, and John Gatesy (2011). "A Supermatrix Analysis of Genomic, Morphological, and Paleontological Data from Crown Cetacea". *BMC Evolutionary Biology* 11.1.

Goerner-Potvin, Patricia and Guillaume Bourque (2018). "Computational Tools to Unmask Transposable Elements". *Nature Reviews Genetics* 19.11, pp. 688–704.

Green, R. E. et al. (2010). "A Draft Sequence of the Neandertal Genome". *Science* 328.5979, pp. 710–722.

Guan, Peiyong and Wing-Kin Sung (2016). "Structural Variation Detection Using Next-Generation Sequencing Data". *Methods* 102, pp. 36–49.

Hahn, Matthew W. and Luay Nakhleh (2016). "Irrational Exuberance for Resolved Species Trees". *Evolution* 70.1, pp. 7–17.

Hailer, F. et al. (2012). "Nuclear Genomic Sequences Reveal That Polar Bears Are an Old and Distinct Bear Lineage". *Science* 336.6079, pp. 344–347.

Hallström, Björn M and A. Janke (2010). "Mammalian Evolution May Not Be Strictly Bifurcating". *Molecular Biology and Evolution* 27.12, pp. 2804–2816.

Hartig, Gerrit et al. (2013). "Retrophylogenomics Place Tarsiers on the Evolutionary Branch of Anthropoids". *Scientific Reports* 3.1.

Hassanin, Alexandre et al. (2012). "Pattern and Timing of Diversification of Cetartiodactyla (Mammalia, Laurasiatheria), as Revealed by a Comprehensive Analysis of Mitochondrial Genomes". *Comptes Rendus Biologies* 335.1, pp. 32–50.

Hattori, Masahira, Satoru Kuhara, Osamu Takenaka, and Yoshiyuki Sakaki (1986). "L1 Family of Repetitive DNA Sequences in Primates May Be Derived from a Sequence Encoding a Reverse Transcriptase-Related Protein". *Nature* 321.6070, pp. 625–628.

Hennig, Willi (1965). "Phylogenetic Systematics". *Annual Review of Entomology* 10.1, pp. 97–116.

Hoban, Sean et al. (2016). "Finding the Genomic Basis of Local Adaptation: Pitfalls, Practical Solutions, and Future Directions". *The American Naturalist* 188.4.

Hormozdiari, F. et al. (2013). "Rates and Patterns of Great Ape Retrotransposition". *Proceedings of the National Academy of Sciences* 110.33, pp. 13457–13462.

Huff, Chad D, Jinchuan Xing, Alan R Rogers, David Witherspoon, and Lynn B Jorde (2010). "Mobile Elements Reveal Small Population Size in the Ancient Ancestors of Homo Sapiens." *Proceedings of the National Academy of Sciences of the United States of America* 107.5, pp. 2147–2152.

Ivancevic, Atma M., R. Daniel Kortschak, Terry Bertozzi, and David L. Adelson (2016). "LINEs between Species: Evolutionary Dynamics of LINE-1 Retrotransposons across the Eukaryotic Tree of Life". *Genome Biology and Evolution* 8.11, pp. 3301–3322.

Jarvis, E. D. et al. (2014). "Whole-Genome Analyses Resolve Early Branches in the Tree of Life of Modern Birds". *Science* 346.6215, pp. 1320–1331.

Jones, Matthew R. et al. (2018). "Adaptive Introgression Underlies Polymorphic Seasonal Camouflage in Snowshoe Hares". *Science* 360.6395, pp. 1355–1358.

Jónsson, Hákon et al. (2014). "Speciation with Gene Flow in Equids despite Extensive Chromosomal Plasticity". *Proceedings of the National Academy of Sciences* 111.52, pp. 18655–18660.

Jurka, J. (1997). "Sequence Patterns Indicate an Enzymatic Involvement in Integration of Mammalian Retroposons". *Proceedings of the National Academy of Sciences* 94.5, pp. 1872–1877.

Jurka, J. et al. (2005). "Repbase Update, a Database of Eukaryotic Repetitive Elements". *Cytogenetic and Genome Research* 110.1-4, pp. 462–467.

Kapitonov, Vladimir V. and Jerzy Jurka (2007). "Helitrons on a Roll: Eukaryotic Rolling-Circle Transposons". *Trends in genetics: TIG* 23.10, pp. 521–529.

Kapitonov, Vladimir V and Jerzy Jurka (2008). "A Universal Classification of Eukaryotic Transposable Elements Implemented in Repbase". *Nature Reviews Genetics* 9, pp. 411–412.

Keane, Michael et al. (2015). "Insights into the Evolution of Longevity from the Bowhead Whale Genome". *Cell Reports* 10.1, pp. 112–122.

Keane, Thomas M., Kim Wong, and David J. Adams (2013). "RetroSeq: Transposable Element Discovery from next-Generation Sequencing Data". *Bioinformatics* 29.3, pp. 389–390.

Krause, Johannes et al. (2008). "Mitochondrial Genomes Reveal an Explosive Radiation of Extinct and Extant Bears near the Miocene-Pliocene Boundary". *BMC Evolutionary Biology* 8.1, p. 220.

Kriegs, Jan Ole et al. (2006). "Retroposed Elements as Archives for the Evolutionary History of Placental Mammals". *PLoS Biology* 4.4, pp. 537–544.

Kumar, Vikas et al. (2017). "The Evolutionary History of Bears Is Characterized by Gene Flow across Species". *Scientific Reports* 7, p. 46487.

Kuramoto, Tae, Hidenori Nishihara, Maiko Watanabe, and Norihiro Okada (2015). "Determining the Position of Storks on the Phylogenetic Tree of Waterbirds by Retroposon Insertion Analysis". *Genome Biology and Evolution* 7.12, pp. 3180–3189.

Kuritzin, Andrej, Tabea Kischka, Jürgen Schmitz, and Gennady Churakov (2016). "Incomplete Lineage Sorting and Hybridization Statistics for Large-Scale Retroposon Insertion Data". *PLOS Computational Biology* 12.3. Ed. by Alon Keinan, e1004812.

Kutschera, Verena E. et al. (2014). "Bears in a Forest of Gene Trees: Phylogenetic Inference Is Complicated by Incomplete Lineage Sorting and Gene Flow". *Molecular Biology and Evolution* 31.8, pp. 2004–2017.

Lammers, Fritjof, Moritz Blumer, Cornelia Rücklé, and Maria A. Nilsson (2019). "Retrophylogenomics in Rorquals Indicate Large Ancestral Population Sizes and a Rapid Radiation". *Mobile DNA* 10.5, pp. 1–9.

Lammers, Fritjof, Susanne Gallus, Axel Janke, and Maria A Nilsson (2017). "Phylogenetic Conflict in Bears Identified by Automated Discovery of Transposable Element Insertions in Low-Coverage Genomes". *Genome Biology and Evolution* 9.10, pp. 2862–2878.

Lander, Eric S et al. (2001). "Initial Sequencing and Analysis of the Human Genome." *Nature* 409.6822, pp. 860–921.

Lankenau, Dirk-Henner and Jean-Nicolas Volff, eds. (2009). *Transposons and the Dynamic Genome*. Genome Dynamics and Stability 4. New York: Springer. 184 pp.

Li, Ruiqiang et al. (2010). "De Novo Assembly of Human Genomes with Massively Parallel Short Read Sequencing". *Genome Research* 20.2, pp. 265–272.

Lin, Ke, Sandra Smit, Guusje Bonnema, Gabino Sanchez-Perez, and Dick de Ridder (2015). "Making the Difference: Integrating Structural Variation Detection Tools". *Briefings in Bioinformatics* 16.5, pp. 852–864.

Liu, Shiping et al. (2014). "Population Genomics Reveal Recent Speciation and Rapid Evolutionary Adaptation in Polar Bears". *Cell* 157.4, pp. 785–794.

Luan, D. D., M. H. Korman, J. L. Jakubczak, and T. H. Eickbush (1993). "Reverse Transcription of R2Bm RNA Is Primed by a Nick at the Chromosomal Target Site: A Mechanism for Non-LTR Retrotransposition". *Cell* 72.4, pp. 595–605.

Ma, Zhanshan, Lianwei Li, Chengxi Ye, Minsheng Peng, and Ya-Ping Zhang (2018). "Hybrid Assembly of Ultra-Long Nanopore Reads Augmented with 10×-Genomics Contigs: Demonstrated with a Human Genome". *Genomics* in press.

Maas, Diede L. et al. (2018). "Rapid Divergence of Mussel Populations despite Incomplete Barriers to Dispersal". *Molecular Ecology* 27.7, pp. 1556–1571.

Maddison, Wayne P. (1997). "Gene Trees in Species Trees". *Systematic Biology* 46.3, p. 523.

Maddison, Wayne P, Michael J Donoghue, and David R Maddison (1984). "Outgroup Analysis and Parsimony". *Systematic Zoology* 33, p. 21.

Mallet, James, Nora Besansky, and Matthew W. Hahn (2016). "How Reticulated Are Species?" *BioEssays* 38.2, pp. 140–149.

Marx, Felix G and R Ewan Fordyce (2015). "Baleen Boom and Bust: A Synthesis of Mysticete Phylogeny, Diversity and Disparity". *Royal Society Open Science* 2 (October), p. 140434.

Mayr, Ernst (1963). *Animal Species and Evolution*. Cambridge, Massachusetts: The Belknap Press of Harvard University Press.

McClintock, B (1950). "The Origin and Behavior of Mutable Loci in Maize". *Proceedings of the National Academy of Sciences of the United States of America* 36.6, pp. 344–355.

McGowen, Michael R., Michelle Spaulding, and John Gatesy (2009). "Divergence Date Estimation and a Comprehensive Molecular Tree of Extant Cetaceans". *Molecular Phylogenetics and Evolution* 53.3, pp. 891–906.

Medvedev, Paul, Monica Stanciu, and Michael Brudno (2009). "Computational Methods for Discovering Structural Variation with Next-Generation Sequencing". *Nature Methods* 6.11, S13–S20.

Meyer, Thomas J. et al. (2012). "An Alu-Based Phylogeny of Gibbons (Hylobatidae)". *Molecular Biology and Evolution* 29.11, pp. 3441–3450.

Mikkelsen, Tarjei S. et al. (2007). "Genome of the Marsupial Monodelphis Domestica Reveals Innovation in Non-Coding Sequences". *Nature* 447.7141, pp. 167–177.

Miller, W. et al. (2012). "Polar and Brown Bear Genomes Reveal Ancient Admixture and Demographic Footprints of Past Climate Change". *Proceedings of the National Academy of Sciences* 109.36, E2382–E2390.

Mirarab, S. et al. (2014). "ASTRAL: Genome-Scale Coalescent-Based Species Tree Estimation". *Bioinformatics* 30.17, pp. i541–i548.

Nellåker, Christoffer et al. (2012). "The Genomic Landscape Shaped by Selection on Transposable Elements across 18 Mouse Strains". *Genome Biology* 13.6, R45.

Nelson, Michael G, Raquel S Linheiro, and Casey M Bergman (2017). "McClintock: An Integrated Pipeline for Detecting Transposable Element Insertions in Whole Genome Shotgun Sequencing Data". *G3: Genes, Genomes, Genetics* 7.8, pp. 2763–2778.

Nikaido, Masato, Oliver Piskurek, and Norihiro Okada (2007). "Toothed Whale Monophyly Reassessed by SINE Insertion Analysis: The Absence of Lineage Sorting Effects Suggests a Small Population of a Common Ancestral Species". *Molecular Phylogenetics and Evolution* 43.1, pp. 216–224.

Nikaido, Masato et al. (2006). "Baleen Whale Phylogeny and a Past Extensive Radiation Event Revealed by SINE Insertion Analysis". *Molecular Biology and Evolution* 23.5, pp. 866–873.

Nilsson, Maria A. et al. (2010). "Tracking Marsupial Evolution Using Archaic Genomic Retroposon Insertions". *PLoS Biology* 8.7. Ed. by David Penny, e1000436.

Nosil, Patrik (2008). "Speciation with Gene Flow Could Be Common". *Molecular Ecology* 17.9, pp. 2103–2106.

Nowak, Ronald M. (1999). *Walker's Mammals of the World*. 6. ed. Baltimore: Johns Hopkins University Press. 837 pp.

O'Donnell, Kathryn A and Kathleen H Burns (2010). "Mobilizing Diversity: Transposable Element Insertions in Genetic Variation and Disease". *Mobile DNA* 1.1, p. 21.

Okada, Norihiro and Mitsuhiro Hamada (1997). "The 3' Ends of tRNA-Derived SINEs Originated from the 3' Ends of LINEs: A New Example from the Bovine Genome". *Journal of Molecular Evolution* 44.S1, S52–S56.

Pagès, Marie et al. (2008). "Combined Analysis of Fourteen Nuclear Genes Refines the Ursidae Phylogeny". *Molecular Phylogenetics and Evolution* 47.1, pp. 73–83.

Pease, James B. and Matthew W. Hahn (2015). "Detection and Polarization of Introgression in a Five-Taxon Phylogeny". *Systematic Biology* 64.4, pp. 651–662.

Pellicciari, C., E. Ronchetti, D. Formenti, R. Stanyon, and M. G. Manfredi Romanini (1990). "Genome Size and «C-Heterochromatic-DNA» in Man and the African Apes". *Human Evolution* 5.3, pp. 261–267.

Peona, Valentina, Matthias H. Weissensteiner, and Alexander Suh (2018). "How Complete Are "Complete" Genome Assemblies?-An Avian Perspective". *Molecular Ecology Resources*, pp. 1–8.

Philippe, Hervé, Frédéric Delsuc, Henner Brinkmann, and Nicolas Lartillot (2005). "Phylogenomics". *Annual Review of Ecology, Evolution, and Systematics* 36.1, pp. 541–562.

Platt, Roy N., Michael W. Vandewege, and David A. Ray (2018). "Mammalian Transposable Elements and Their Impacts on Genome Evolution". *Chromosome Research* 26.1-2, pp. 25–43.

Platt, Roy N. et al. (2015). "Targeted Capture of Phylogenetically Informative Ves SINE Insertions in Genus Myotis". *Genome Biology and Evolution* 7.6, pp. 1664–1675.

Prado-Martinez, Javier et al. (2013). "Great Ape Genetic Diversity and Population History". *Nature* 499.7459, pp. 471–475.

Quinlan, Aaron R. and Ira M. Hall (2010). "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features". *Bioinformatics* 26.6, pp. 841–842.

Ray, David A., Jinchuan Xing, Abdel-Halim Salem, and Mark A. Batzer (2006). "SINEs of a Nearly Perfect Character". *Systematic Biology* 55.6, pp. 928–935.

Rocha Jr., Robert C., Phillip J. Clapham, and Yulia Ivashchenko (2015). "Emptying the Oceans: A Summary of Industrial Whaling Catches in the 20th Century". *Marine Fisheries Review* 76.4, pp. 37–48.

Roman, J. (2003). "Whales Before Whaling in the North Atlantic". *Science* 301.5632, pp. 508–510.

Rychel, Amanda L, Tod W Reeder, and Annalisa Berta (2004). "Phylogeny of Mysticete Whales Based on Mitochondrial and Nuclear Data". *Molecular Phylogenetics and Evolution* 32.3, pp. 892–901.

Saitou, N. and M. Nei (1987). "The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees." *Molecular Biology and Evolution* 4.4, pp. 406–425.

Sasaki, Takeshi et al. (2005). "Mitochondrial Phylogenetics and Evolution of Mysticete Whales". *Systematic biology* 54.1, pp. 77–90.

Shedlock, A M, M C Milinkovitch, and N Okada (2000). "SINE Evolution, Missing Data, and the Origin of Whales." *Systematic biology* 49.4, pp. 808–817.

Shedlock, Andrew M., Kazuhiko Takahashi, and Norihiro Okada (2004). "SINEs of Speciation: Tracking Lineages with Retroposons". *Trends in Ecology & Evolution* 19.10, pp. 545–553.

Shimamura, Mitsuru, Hideaki Abe, Masato Nikaido, Kazuhiko Ohshima, and Norihiro Okada (1999). "Genealogy of Families of SINEs in Cetaceans and Artiodactyls: The Presence of a Huge Superfamily of tRNA(Glu)-Derived Families of SINEs". *Molecular Biology and Evolution* 16.8, pp. 1046–1060.

Shimamura, Mitsuru et al. (1997). "Molecular Evidence from Retroposons That Whales Form a Clade within Even-Toed Ungulates". *Nature* 388.6643, pp. 666–670.

Slatkin, Montgomery and Fernando Racimo (2016). "Ancient DNA and Human History". *Proceedings of the National Academy of Sciences*, p. 201524306.

Smit, Arian FA (1999). "Interspersed Repeats and Other Mementos of Transposable Elements in Mammalian Genomes". *Current Opinion in Genetics & Development* 9.6, pp. 657–663.

Solís-Lemus, Claudia, Paul Bastide, and Cécile Ané (2017). "PhyloNetworks: A Package for Phylogenetic Networks". *Molecular Biology and Evolution* 34.12, pp. 3292–3298.

Sotero-Caio, Cibele G, Roy N Platt, Alexander Suh, and David A Ray (2017). "Evolution and Diversity of Transposable Elements in Vertebrate Genomes". *Genome Biology and Evolution* 9.1, pp. 161–177.

Spilliaert, R. et al. (1991). "Species Hybridization between a Female Blue Whale (Balaenoptera Musculus) and a Male Fin Whale (B. Physalus): Molecular and Morphological Documentation". *The Journal of Heredity* 82.4, pp. 269–274.

Steeman, Mette E. et al. (2009). "Radiation of Extant Cetaceans Driven by Restructuring of the Oceans". *Systematic Biology* 58.6, pp. 573–585.

Stewart, Chip et al. (2011). "A Comprehensive Map of Mobile Element Insertion Polymorphisms in Humans". *PLoS Genetics* 7.8. Ed. by Harmit S. Malik, e1002236.

Sudmant, P. H. et al. (2015). "An Integrated Map of Structural Variation in 2,504 Human Genomes". *Nature* 526.7571, pp. 75–81.

Suh, Alexander, Linnéa Smeds, and Hans Ellegren (2015). "The Dynamics of Incomplete Lineage Sorting across the Ancient Adaptive Radiation of Neoavian Birds". *PLOS Biology* 13.8. Ed. by David Penny, e1002224.

Suh, Alexander, Linnéa Smeds, and Hans Ellegren (2017). "Abundant Recent Activity of Retrovirus-like Retrotransposons within and among Flycatcher Species Implies a Rich Source of Structural Variation in Songbird Genomes". *Molecular Ecology* (July), pp. 1–13.

Suh, Alexander et al. (2011). "Mesozoic Retroposons Reveal Parrots as the Closest Living Relatives of Passerine Birds". *Nature Communications* 2.443, pp. 1–7.

The 1000 Genomes Project Consortium et al. (2015). "A Global Reference for Human Genetic Variation". *Nature* 526.7571, pp. 68–74.

Thung, Djie Tjwan et al. (2014). "Mobster: Accurate Detection of Mobile Element Insertions in next Generation Sequencing Data". *Genome Biology* 15.448, p. 11.

Treangen, Todd J. and Steven L. Salzberg (2012). "Repetitive DNA and Next-Generation Sequencing: Computational Challenges and Solutions". *Nature Reviews Genetics* 13.1, pp. 36–46.

Trizzino, Marco et al. (2017). "Transposable Elements Are the Primary Source of Novelty in Primate Gene Regulation". *Genome Research* 27.10, pp. 1623–1633.

Tsagkogeorga, Georgia, Joe Parker, Elia Stupka, James A. Cotton, and Stephen J. Rossiter (2013). "Phylogenomic Analyses Elucidate the Evolutionary Relationships of Bats". *Current Biology* 23.22, pp. 2262–2267.

Tuzun, Eray et al. (2005). "Fine-Scale Structural Variation of the Human Genome". *Nature Genetics* 37.7, pp. 727–732.

Van de Lagemaat, L. N. (2005). "Genomic Deletions and Precise Removal of Transposable Elements Mediated by Short Identical DNA Segments in Primates". *Genome Research* 15.9, pp. 1243–1249.

Van der Auwera, Geraldine A. et al. (2013). "From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline: The Genome Analysis Toolkit Best Practices Pipeline". *Current Protocols in Bioinformatics*. Ed. by Alex Bateman, William R. Pearson, Lincoln D. Stein, Gary D. Stormo, and John R. Yates. Hoboken, NJ, USA: John Wiley & Sons, Inc., pp. 11.10.1–11.10.33.

Van't Hof, Arjen E. et al. (2016). "The Industrial Melanism Mutation in British Peppered Moths Is a Transposable Element". *Nature* 534.7605, pp. 102–105.

Vassetzky, Nikita S. and Dmitri A. Kramerov (2002). "CAN—a Pan-Carnivore SINE Family". *Mammalian Genome* 13.1, pp. 50–57.

Walters-Conte, K. B., D. L. E. Johnson, M. W. Allard, and J. Pecon-Slattery (2011). "Carnivore-Specific SINEs (Can-SINEs): Distribution, Evolution, and Genomic Impact". *Journal of Heredity* 102.S1, S2–S10.

Williams, David M. and Peter L. Forey (2004). *Milestones in Systematics*. Systematics Association Special Volume. CRC Press. 309 pp.

Witherspoon, D. J. et al. (2013). "Mobile Element Scanning (ME-Scan) Identifies Thousands of Novel Alu Insertions in Diverse Human Populations". *Genome Research* 23.7, pp. 1170–1181.

Wong, Kim, Thomas M Keane, James Stalker, and David J Adams (2010). "Enhanced Structural Variant and Breakpoint Detection Using SVMerge by Integration of Multiple Detection Methods and Local Assembly". *Genome Biology* 11.12, R128.

Yang, Ziheng and Bruce Rannala (2012). "Molecular Phylogenetics: Principles and Practice". *Nature Reviews Genetics* 13.5, pp. 303–314.

Ye, K., M. H. Schulz, Q. Long, R. Apweiler, and Z. Ning (2009). "Pindel: A Pattern Growth Approach to Detect Break Points of Large Deletions and Medium Sized Insertions from Paired-End Short Reads". *Bioinformatics* 25.21, pp. 2865–2871.

Yim, Hyung-Soon et al. (2014). "Minke Whale Genome and Aquatic Adaptation in Cetaceans." *Nature genetics* 46.1, pp. 88–92.

Yu, Yun, Jianrong Dong, Kevin J. Liu, and Luay Nakhleh (2014). "Maximum Likelihood Inference of Reticulate Evolutionary Histories". *Proceedings of the National Academy of Sciences* 111.46, pp. 16448–16453.

Publications

4.1 Phylogenetic conflict in bears identified by automated discovery of transposable element insertions in low-coverage genomes

Fritjof Lammers, Susanne Gallus, Axel Janke & Maria A. Nilsson

Published in
Genome Biology and Evolution 9 (10): 2862–78 (2017).
doi:10.1093/gbe/evx170

Erklärung zu den Autorenanteilen an der Publikation

Was hat der Promovierende bzw. was haben die Koautoren beigetragen?

(1) zu Entwicklung und Planung
FL 70 %, MAN 20 %, AJ 10 %

(2) zur Durchführung der einzelnen Untersuchungen und Experimente
Entwicklung Bioinformatik, Programmierung: FL (40 %); Genomanalyse: FL (40 %); Primerdesign: FL, SG (10%); DNA Isolation, PCR, Sequenzierung: SG, MAN (10%)

(3) zur Erstellung der Datensammlung und Abbildungen
Abbildungen, Tabellen: FL (90%); Sequenzbearbeitung der Labordaten: SG (10%)

(4) zur Analyse und Interpretation der Daten
Datenanalyse Bioinformatik: FL (80%); Zusammenfassende Interpretation: FL, MAN, AJ (20%)

(5) zum Verfassen des Manuskripts
FL 70 %, MAN: 20 %, AJ: 10 %

Zustimmende Bestätigungen der oben genannten Angaben:

_____ , _____	
Datum, Ort	Unterschrift Promovierender
_____ , _____	
Datum, Ort	Unterschrift Betreuer
_____ , _____	
Datum, Ort	ggfs. Unterschrift *corresponding author*

GBE

Phylogenetic Conflict in Bears Identified by Automated Discovery of Transposable Element Insertions in Low-Coverage Genomes

Fritjof Lammers[1,2], Susanne Gallus[1], Axel Janke[1,2], and Maria A. Nilsson[1,*]

[1]Senckenberg Biodiversity and Climate Research Centre, Senckenberg Gesellschaft für Naturforschung, Frankfurt am Main, Germany

[2]Institute for Ecology, Evolution & Diversity, Biologicum, Goethe University Frankfurt, Frankfurt am Main, Germany

*Corresponding author: E-mail: maria.nilsson-janke@senckenberg.de.

Accepted: August 28, 2017

Abstract

Phylogenetic reconstruction from transposable elements (TEs) offers an additional perspective to study evolutionary processes. However, detecting phylogenetically informative TE insertions requires tedious experimental work, limiting the power of phylogenetic inference. Here, we analyzed the genomes of seven bear species using high-throughput sequencing data to detect thousands of TE insertions. The newly developed pipeline for TE detection called TeddyPi (TE detection and discovery for Phylogenetic Inference) identified 150,513 high-quality TE insertions in the genomes of ursine and tremarctine bears. By integrating different TE insertion callers and using a stringent filtering approach, the TeddyPi pipeline produced highly reliable TE insertion calls, which were confirmed by extensive in vitro validation experiments. Analysis of single nucleotide substitutions in the flanking regions of the TEs shows that these substitutions correlate with the phylogenetic signal from the TE insertions. Our phylogenomic analyses show that TEs are a major driver of genomic variation in bears and enabled phylogenetic reconstruction of a well-resolved species tree, despite strong signals for incomplete lineage sorting and introgression. The analyses show that the Asiatic black, sun, and sloth bear form a monophyletic clade, in which phylogenetic incongruence originates from incomplete lineage sorting. TeddyPi is open source and can be adapted to various TE and structural variation callers. The pipeline makes it possible to confidently extract thousands of TE insertions even from low-coverage genomes (~10×) of nonmodel organisms. This opens new possibilities for biologists to study phylogenies and evolutionary processes as well as rates and patterns of (retro-)transposition and structural variation.

Key words: retrotransposition, bears, Ursidae, phylogeny, evolution, transposable elements.

Introduction

In an innovative study almost 20 years ago, rare genomic changes were used to confirm the close relationship between hippopotamus (Artiodactyla) and whales (Cetacea) (Shimamura et al. 1997; Nikaido et al. 1999). Transposable element (TE) insertions are a type of rare genomic changes that propagate in the genome via copy-and-paste (retrotransposons) or cut-and-paste (DNA transposons) mechanisms. Germline transposition events are passed on to descendants, making it possible to deduce their phylogenetic relationships (Shimamura et al. 1997; Nikaido et al. 1999). In contrast to nucleotide substitutions which are prone to homoplasy by parallelisms, convergence, and reversals, TE insertions are virtually homoplasy free. Parallel integration of TE insertions in the same loci in different species is highly improbable due to low-germline insertion rates and the presence of different active TE families (Ray et al. 2006). Finally, the exact removal of TE insertions is very rare and usually leaves a detectable genetic "scar" (van de Lagemaat et al. 2005). These features are very valuable for the understanding of deep or complex divergences, like the early radiation of mammals and birds (Churakov et al. 2009; Nishihara et al. 2009; Hallström and Janke 2010; Suh et al. 2015).

Detecting phylogenetically informative TE insertions was initially challenging because fully sequenced genomes were not available (Shimamura et al. 1997; Nikaido et al. 1999). Therefore, only experimental work could identify candidate TE loci of which often only a minor fraction were phylogenetically informative (Shimamura et al. 1997; Nikaido et al. 1999). Although, the increasing availability of genome assemblies

© The Author 2017. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

and new methods have allowed computational identification of phylogenetically informative TE insertions, extending the taxon sampling for species without available genomes relied still on extensive experimental testing (Kriegs et al. 2006; Churakov et al. 2009). Alternative methods, not based on genome assemblies can only identify a limited number of informative TE insertions (Suh et al. 2012; Kuramoto et al. 2015). Finally, experimental enrichment protocols for TE insertions can identify thousands of informative loci, but require knowledge of the TE sequence and are biased toward loci with existing TE insertions (Platt et al. 2015). A recently developed bioinformatic approach to detect novel TE insertions uses the information from discordantly mapped paired-end short reads without requiring a de novo genome assembly for each species (Medvedev et al. 2009). Such "TE calling" methods have allowed biologists to study TE insertion dynamics and other structural variations (SV) on a population-scale (Hormozdiari et al. 2013; Sudmant et al. 2015). This approach has been successfully applied to the great apes and to mice (Nellåker et al. 2012; Hormozdiari et al. 2013) showing its potential for phylogenetic inference. However, as yet, no phylogenetic study has applied TE calling methods to nonmodel organisms, for which often only draft genome assemblies and low-coverage resequencing data are available.

A long-standing question in phylogenetics is determining the evolutionary history of bears (Ursidae), for which different scenarios have been reconstructed from analyses of mitochondrial, autosomal, and gonosomal DNA sequences. In particular, the six ursine species that include the polar (*Ursus maritimus*) and brown bear (*Ursus arctos*), share a complex evolutionary history due to their rapid radiation during the Pliocene (5–3.5 Million years ago (Ma)) (Kumar et al. 2017). The best studied examples are polar bears, which according to mitochondrial DNA (mtDNA) are nested within the brown bears (Yu et al. 2007). However, analyses of nuclear DNA showed that polar bears are genetically distinct and the sistergroup to brown bears (Hailer et al. 2012). The phylogeny of the American black bear (*Ursus americanus*) and the three South-East Asian bear species is less understood with deviating mtDNA and nuclear gene trees (Yu et al. 2007; Pagès et al. 2008; Kutschera et al. 2014). Phylogenomic analyses reconstructed the American black bear as the sister group to a monophyletic polar and brown bear lineage and show that the three South-East Asian bears form a clade with the Asiatic black bear (*Ursus thibetanus*) being the sister group to sun (*Ursus malayanus*) and sloth bear (*Ursus ursinus*) (Kutschera et al. 2014; Kumar et al. 2017).

The observed phylogenetic incongruence among bears can be caused by introgressive hybridization and incomplete lineage sorting (ILS) (Maddison 1997). As such, the analysis of genome-wide data is necessary to understand these complex processes (Delsuc et al. 2005). However, the lack of whole genome sequences inhibited efficient screening for phylogenetically informative TE insertion events until the polar bear genome sequence and genome data of all other bear species became available (Miller et al. 2012; Liu et al. 2014; Kumar et al. 2017). These new genome data have allowed us to detect TE insertions as additional independent phylogenomic markers to study the evolution of Ursidae. We developed the TeddyPi (TE detection and discovery for phylogenetic inference) pipeline to process data from TE and SV callers. TeddyPi pursues the idea of integrating different TE callers (Lin et al. 2015; Nelson et al. 2017) and extends it to routinely integrate TE insertion data sets from multiple samples to track integrations of TEs in orthologous loci and to create presence/absence tables for phylogenetic inference.

We applied TeddyPi to whole-genome sequencing data of all ursine bears and the monotypic subfamily Tremarctinae to extract phylogenetically informative markers that are independent from nucleotide substitution analyses. We aimed to study the evolutionary history of bears and test whether TE insertions identify the same signals of gene flow and ILS as in a previous nucleotide-based study (Kumar et al. 2017). This recently generated genome data of all ursine bears made it possible to observe nucleotide substitutions in the flanks around the TE insertions, that are mutationally saturated for deeper divergences. To validate the in silico TE calls made by TeddyPi, 151 loci were validated experimentally by PCR and Sanger sequencing. The TeddyPi pipeline extracted an extensive catalog of 150,513 TE insertions to reconstruct the first TE-derived species tree of bears and revealed varying rates of TE accumulation in their genomes.

Materials and Methods

The TeddyPi Pipeline

The TeddyPi pipeline is a modular framework to process TE and SV calls and to prepare data sets for phylogenetic inference. The application is written in Python and utilizes established code libraries for biological computing. Parameters and the filter pipeline are configured with comprehensively structured configuration files and allow the user to create tailor-made filtering pipelines for a variety of variant callers. The first module (teddypi.py) processes each sample genome individually and filters the output of the selected variant callers. Several filters and merge-functions are included in this module, and a flexible codebase allows implementation of new functions with little programming knowledge. In the same module, large deletions are transformed to reference-insertion calls on the basis of annotated TEs in the reference genome. It is also possible to make intersections or create nonredundant data sets of the input data in this step. In the second module (tpi_ortho.py), TE insertion data is combined across a set of samples (typically different taxa) to generate presence/absence matrices for reference insertions (Ref+) and nonreference insertions (Ref−) separately. Finally, the last module (tpi_unite.py), merges both matrices to a

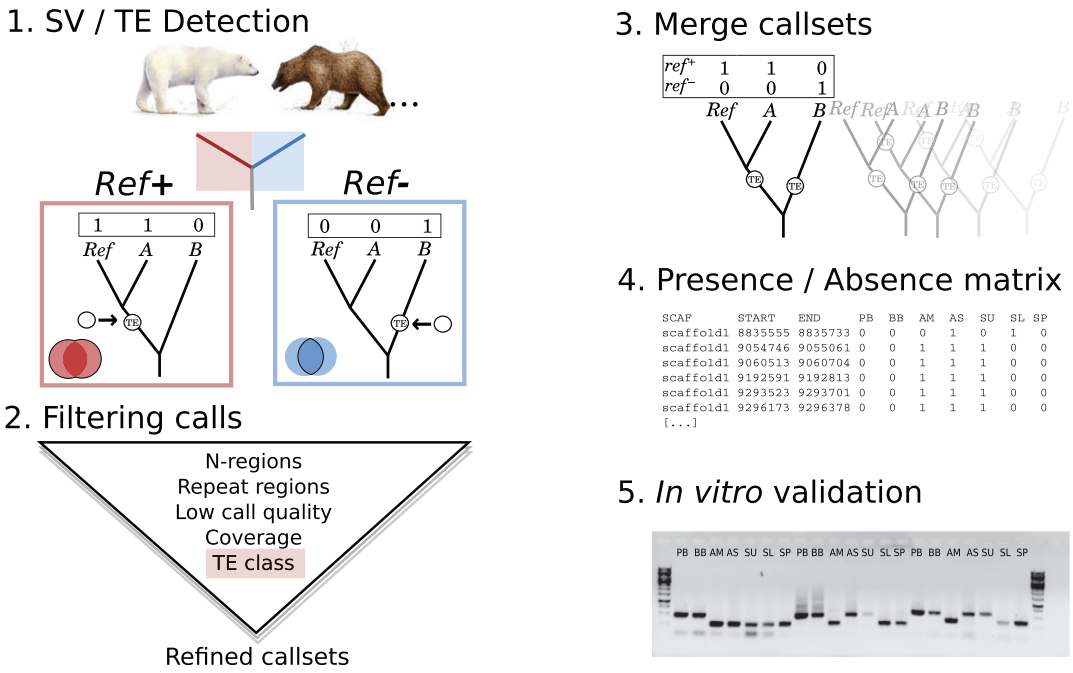


Fig. 1.—Schematic illustration of the TeddyPi pipeline. (1) Transposable Element (TE) and Structural Variation (SV) callers detect reference (Ref+, red) and nonreference (Ref−, blue) TE insertions from reads mapped to a reference genome. The boxed trees show a schematic phylogeny with the reference genome (Ref) and two other taxa (A and B). The TE insertion is shown by an arrow and indicates Ref+ and Ref− detection depending on which branch the TE inserted. (2) TE calls are filtered based on the polar bear genome annotation, call quality, and sequencing coverage across the genome. Different TE classes are collected separately. (3) Sets of TE calls (call sets) for each individual genome are merged to create a comprehensive presence/absence matrix (4) that is used for phylogenetic inference and (5) to select loci for in vitro validation.

comprehensive presence/absence matrix that can be exported in tabular-text and NEXUS format. A schematic overview is shown in figure 1 and a flowchart of the pipeline in supplementary figure 1, Supplementary Material online. TeddyPi is open source and can be accessed on https://github.com/mobilegenome/teddypi, last accessed September 2017. Easy configuration and the modular architecture make it convenient to adapt TeddyPi to process data from a broad range of TE/SV callers or other integration pipelines such as SVMerge or McClintock (Wong et al. 2010; Nelson et al. 2017). TeddyPi can be applied to any group of organisms where accurate TE/SV calling is feasible.

Taxon Sampling and Genome Sequencing

Whole genome sequencing data generated with Illumina HiSeq technology from Kumar et al. (2017) and Miller et al. (2012) for six ursine bear species and the spectacled bear (*Tremarctos ornatus*) were obtained. For mapping, paired-end reads (100–125 base pairs (bp) long) were quality-trimmed with Trimmomatic (Bolger et al. 2014), mapped with BWA (Li and Durbin 2010), and duplicated reads were marked. In total, nine genomes with a mean coverage of 13.7× from seven species were analyzed (supplementary table 1, Supplementary Material online). In comparison to the giant panda (*Ailuropoda melanoleuca*) genome sequence (ailMel1), the polar bear genome sequence (Liu et al. 2014) has higher contiguity and contains potentially better-assembled repeats because it is based on longer sequencing reads. Therefore, the polar bear was the preferred choice for reference mapping.

Considerations for Nested Reference Genomes

Programs to detect TE insertions (in analogy to SNP callers named TE callers) depend on a pairwise comparison between the paired-end short reads of a sample and the reference genome the reads were mapped to. As most published TE callers can only detect nonreference (Ref−) TE insertions it is beneficial to have a reference genome that is phylogenetically placed as the outgroup to the taxa under study to detect insertions across the complete phylogeny (supplementary fig. 2, Supplementary Material online). If this is not possible,

the use of only nonreference TE callers will lead to unresolved internodes and a skewed phylogenetic interpretation. For example, when the reference genome is nested inside the ingroup/tree, only TE insertions on the terminal branches are detectable and certain internodes cannot be resolved (supplementary fig. 2, Supplementary Material online). To overcome such a bias, reference (Ref+) TE insertions (i.e., those shared with the reference genome) need to be considered, too. Ref+ TE insertions can not be called directly, but are inferred from deletion calls made from the mapped short read data. These deletions resemble insertions in the reference genome sequence. From the reference genome, the sequence of the insertion can be extracted and screened for similarity to known TEs. TeddyPi inverts deletion calls that intersect with TE sequences in the reference genome to infer Ref+ TE insertions (supplementary fig. 3, Supplementary Material online).

Analysis of TEs in the Polar Bear Genome Sequence

Repetitive elements in the polar bear genome were identified using RepeatMasker in sensitive mode (-s) searching for carnivore-specific repeats (Repbase version 20140131). The script createRepeatLandscape.pl provided with RepeatMasker was used to calculate the repeat landscape. We explored the diversity of LINE1 copies in the polar bear genome to find active copies that can drive retrotransposition or inactive copies incapable of retrotransposition, for example, by the presence of premature stop codons in the open reading frame (ORF) 2 of the LINE1. The LINE1 ORF2 sequence was retrieved from a full-length LINE1 found on the polar bear Y chromosome (Bidon et al. 2015) and used as a BLAST query against the polar bear genome sequence (Altschul et al. 1990). Hits were filtered for full length, coding ORF2 copies and a maximum of three mismatches. Then, these sequences plus 7,000-bp flanking sequence on 5′ and 3′ ends were extracted from the polar bear genome sequence. Within these sequences, a BLAST search for a coding LINE1 ORF1 sequence was performed to find LINE1 copies containing two coding ORFs. As an additional proxy for LINE1 activity, we screened the polar bear and giant panda genome for the U6 snRNA (Accession No: M14486.1) using BLAST. According to Doucet et al. (2015), all hits with >97.5% identity, 26-bp alignment length and an E-value of < 10 were considered as full-length hits. Additionally, we annotated 146,268 gaps totaling to 38 mega base pairs (Mb) in the polar bear genome; the majority of these gaps (138,041) were >1 bp.

Detection of Nonreference (Ref−) TE Insertions

Reference mapped short reads were processed with RetroSeq (Keane et al. 2013) and Mobster (Thung et al. 2014) to identify insertions that are present in the corresponding genome while being absent in the reference genome. For RetroSeq, a minimum mapping quality of 30 and a TE mapping identity of 90% at 50% length were used. The upper coverage threshold was set to 2.5× of the samples' sequencing depth. Mobster was run with default settings. A library of 593 carnivore specific TE sequences was retrieved from Repbase (Jurka et al. 2005), and used as consensus database for both programs. Mobster and RetroSeq used this database to identify reads that match the consensus TE sequence and thereby inferred the type of TE that has been integrated. In addition, RetroSeq identifies reads matching the RepeatMasker track of the reference genome. Using the TeddyPi pipeline, callsets (i.e., the sets of called TE variants) from RetroSeq and Mobster were filtered for calls falling within regions of undetermined bases (N) plus a window of 200 bp in the polar bear genome. Calls were also filtered, if they were supported by less than five reads or when located within 100 bp of annotated TEs of the same type in the polar bear genome. For stringency, both data sets were masked for regions that had a depth of coverage <33% or >250% of the mean sample coverage. Thereby, regions of ambiguously mapped reads or with insufficient coverage for TE calling were excluded. Only overlapping calls from both programs were further processed.

Detection of Reference Insertion (Ref+) TE Insertions

To detect TE insertions absent from at least one of the low-coverage bear genomes and present in polar bear reference genome (Ref+ insertions), Pindel (Ye et al. 2009), and Breakdancer (Chen et al. 2009) mined the genomes for deletions, that are indicative of insertions in the reference genome (Nellåker et al. 2012). Pindel uses split-read (SR) information to obtain breakpoint information at a single-nucleotide level resolution and was run with the following parameters –report_interchromosomal_events false, –anchor_quality 30, -w 40. Only deletions were considered for further processing. BreakDancer was run using a maximum variant size of 10 kilo bases (kb) and requiring at least five supporting reads to make a SV call. BreakDancer utilizes only discordant reads and does not utilize SRs for SV-calling. Therefore, start- and end-coordinates from the deletions were used. For each sample, book-ended (i.e., those directly after another) calls and overlapping calls were merged, filtered for N-regions (+200 bp flanking sequence) and tandem repeats (+50 bp) in the reference genome. All calls in regions with a depth of coverage <33% or >250% of the average were excluded. The calls from Pindel and Breakdancer were merged to a nonredundant set. The start/end coordinates or if available, the breakpoint of the deletion plus a window of ± 50 bp were used to detect intersections with annotated repeats in the polar bear reference genome. Deletion calls that matched duplicate RepeatMasker hits and appeared twice, were merged. When coordinates overlapped with more than one TE in the reference genome, and one was a recent SINE insertion (i.e., SINEC_Ame subfamily) while the other TE(s) were not known to be active within Carnivora, it was called as "SINE derived". If coordinates overlapped with

different types of annotated TEs, and more than one was potentially active, the event was recorded as "complex". Predicted deletion loci of more than one sample were attributed to the same locus if both were intersecting with the reference TE and the distance was <100 bp. To obtain reference insertion (Ref+) calls, presence/absence information was inverted (supplementary fig. 2, Supplementary Material online).

Integration of Ref+ and Ref− Call Sets, Filtering, and Processing

To combine the insertion and deletion data sets, results were integrated across all species. This module of TeddyPi (tpi_ortho.py) loads the final call sets for all species, sorts these by position, and merges overlapping and book-ended calls if not done before. Then, BedTools window is called via pybedtools to create a presence/absence matrix (coded as 1 and 0, respectively) over all variants and taxa (variant × taxa) (Quinlan and Hall 2010; Dale et al. 2011).

Despite originating from the same insertion event, breakpoint estimates might differ slightly between taxa. Therefore, overlapping, book-ended, and events within 100 bp of each other were merged using BedTools. Presence/absence information from deletion calls was inverted (1 ↔ 0) to obtain reference insertions (Ref+) calls. The state of TE insertions in the reference genome was added with either 1 or 0 for Ref+ and Ref− events, respectively. Callsets for Ref+ and Ref− were saved as a tab-separated file and converted to a NEXUS character matrix using the python-nexus package (Greenhill S. unpublished).

Merging Ref+ and Ref− Callsets, and Correcting for Missing Data

Ref+ and Ref− data sets were merged in the tpi_unite.py module of TeddyPi and a final presence/absence matrix was created. A synthetic outgroup with state "0" for all loci was added. For the Ref− data set, loci were coded as missing data ("?" in the NEXUS matrix) for samples with an insufficient or excessive depth of coverage. The criteria were set for each sample individually to include only loci with coverage between 0.33× and 2.5× of the samples mean coverage.

Phylogenetic Inference from TE Insertion Calls

We processed SINE and LINE1 callsets separately and created Dollo parsimony trees in PAUP* (Swofford 2002) using the heuristic search with 500 replicates. Bootstrap support was calculated from 1,000 replicates. The trees were rooted using the synthetic outgroup. The number of SINE insertions for species-tree congruent and alternative topologies were obtained from the presence/absence matrices and analyzed using the KKSC test that conceptually transfers the *D*-statistics to TE insertion data (Durand et al. 2011; Kuritzin et al. 2016). The KKSC test evaluates the number of phylogenetically conflicting TE insertions among three taxa and uses binomial distribution to test for the probability of hybridization or ILS as cause of the observed insertion pattern.

Median networks for SINE insertions were calculated in SplitsTree 4 (Huson and Bryant 2006). Phylogenetic networks for Ref+ and Ref− data were calculated separately using all SINEs and LINE1s.

Estimating TE Insertion Rates

SINE and LINE insertion counts were extracted from the parsimony-tree branch lengths and were divided by the divergence times (in Myr) estimated previously (Kumar et al. 2017) to get estimates on the relative insertion rate. To estimate per-generation insertion rates, the generation time for the polar and brown bear was assumed to be 10 years (Tallmon et al. 2004; Cronin et al. 2009) and 6 years for the other bear species (Onorato et al. 2004; Kutschera et al. 2014).

Genomic Context of TE Insertions

The genomic context of the TE insertions was evaluated using the genome annotation from the polar bear genome (Liu et al. 2014). The TE insertion catalog was screened for overlaps with 3′- and 5′-UTRs, introns, exons, and intergenic regions.

Flanking Sequence Analysis of TE Insertion Loci

TE insertions and the substitutions in their flanking genomic regions are expected to share the same evolutionary history. We sought to explore the congruence in phylogenetic signal between TEs and flanking regions and to determine the range of the phylogenetic congruence (i.e., the spatial extent in bp) between them. To this end, consensus sequence alignments were created using substitution calls from Kumar et al. (2017). First, 10-kb sequence up- and downstream of the insertion site were extracted and the maximum likelihood (ML) phylogeny was inferred with RaxML (Stamatakis 2014) for each flank as well as for the concatenated sequence of both flanks. For automation and calling RAxML, the Dendropy package was utilized (Sukumaran and Holder 2010). To account for possibly misaligned reads around the insertion site, the first 500 bp on each side of the insertion site were excluded. The question was, whether the flanking sequence yields the same phylogenetic signal as the presence/absence pattern of the TE insertion. Therefore, we checked if the species carrying the TE insertion form a monophyletic group in the ML-trees using the ETE toolkit (Huerta-Cepas et al. 2016). Furthermore, to gain insight in the phylogenetic signal in the TE flanking region a sliding window approach was applied to the same 10-kb flanking regions using nonoverlapping 1-kb windows. For each window, sites were counted showing the same

phylogenetic signal as the TE insertion and then divided by the number of segregating sites.

Experimental Validation Screening

From the in silico data set, loci were randomly selected for experimental verification. DNA samples from all ursine bears and the spectacled bear were included. For the Asian bear species and the spectacled bear, the same DNA samples were used for validation as for the Illumina genome sequencing. We selected loci containing TE insertions supporting different topologies (supplementary table 2, Supplementary Material online) including topologies in conflict with the species tree (e.g., presence in American black and Asiatic black bear or American black bear and sun bear).

For primer design, consensus sequence alignments that spanned 4-kb up- and downstream of the predicted TE insertion site were extracted from Kumar et al. (2017). PCR primers were generated with primer3 to be located ~200 bp from the TE insertion site (Untergasser et al. 2012). Primers are listed in the supplementary data 1, Supplementary Material online. Each locus was amplified using 8 ng of DNA per species and Amplicon Taq (VWR) in a touchdown PCR. Banding patterns were examined using gel-electrophoresis agarose gels along with a DNA marker (ThermoFisher GeneRuler 1Kb). The fragment length of each PCR product was estimated and species that had the indication of a TE insertion were recorded. The PCR amplicons were Sanger-sequenced in both directions using the ABI 3730 DNA Analyzer. The type of the inserted sequence was determined by querying the sequence against Repbase (Jurka et al. 2005) (www.girinst.org; last accessed September 2017). For 13 markers, PCR products were sequenced of all or nearly all bear species to verify the phylogenetic information of the loci. The alignments were screened for the TE type, the orientation, target site duplications (TSDs), and the integrity of the flank. Two markers were specifically selected and sequenced to investigate the absence of a SINEC1_Ame in the polar bear (marker 40 and 122).

Experimentally confirmed insertion patterns were compared with the computationally predicted insertions at the same locus. We considered each matching insertion status (predicted: absence—PCR: absence/predicted: presence—PCR: presence) as correctly called. If the PCR product indicated presence of a TE insertion but no TE call was made, the locus was recorded as false negative (FN) and false positive (FP) for the opposite case. If a PCR reaction did not yield an amplicon for a locus, the locus was flagged as inconclusive.

Results

Transposable Elements in Ursine Bears

Our screening of the interspersed repeats in the polar bear reference genome identified 1,223,168 SINEs (8.4% of the genome), 978,888 LINEs (21.3%), 320,346 LTR retrotransposons (5.3%), as well as 340,447 DNA transposons (3.1%) (supplementary table 3, Supplementary Material online). In total, the polar bear genome comprised 38.1% interspersed repeats, similar to other carnivores like the giant panda, dog, or cat (Lindblad-Toh et al. 2005; Pontius et al. 2007; Li et al. 2010). The most abundant and recently active SINE-family in carnivore genomes is the lysine-tRNA derived SINEC (Walters-Conte et al. 2011). In Ursidae, SINEC1_AMe is the most frequent SINE subfamily in both the polar bear and giant panda genomes with 249,740 copies and 237,604 copies, respectively. SINEC1_Ame has a consensus length of 201 bp and was initially described from the giant panda genome (Li et al. 2010). SINEC elements are thought to be LINE1 propagated, and a screen for potentially active full-length LINE1s revealed 535 copies with two intact ORFs in the polar bear genome. The U6 snRNA that has been strongly associated with LINE1 activity in mammalian genomes (Doucet et al. 2015) was found in 67 copies in the polar bear genome sequence. Repeat landscapes of both the polar bear and giant panda genomes indicate the presence of low divergent and thus recently active SINEs (supplementary fig. 4, Supplementary Material online).

Detecting Ref− Insertions

In all analyzed samples, the programs RetroSeq (Keane et al. 2013) and Mobster (Thung et al. 2014) found 696,041 and 491,193 Ref− TE insertions, respectively (supplementary tables 4 and 5, Supplementary Material online). Despite the difference in numbers of raw calls, the number of SINEs and LINEs selected from the unfiltered data sets of RetroSeq and Mobster are very similar (~300,000 SINEs, ~135,000 LINEs). Still, data sets from both programs differed in susceptibility to the subsequent filtering pipeline, indicating differences in the overall call-quality (supplementary tables 4–7, Supplementary Material online). Thus, after filtering, 50% more SINEs were obtained from Mobster than from RetroSeq. For LINEs, 25% more calls from RetroSeq were retained by TeddyPi (supplementary table 8, Supplementary Material online). After merging data from RetroSeq and Mobster, the final Ref− insertion data set consisted of 84,462 SINEs and 7,734 LINEs (supplementary table 8, Supplementary Material online).

Detecting Ref+ Insertions

A different approach was necessary to identify Ref+ TE insertions due to nested position of the polar bear in the ursine species tree (supplementary fig. 2, Supplementary Material online). The two SV callers Pindel (Ye et al. 2009) and BreakDancer (Chen et al. 2009) identified 10,527,959 deletions in the nine bear genomes. Of these ~10.5 million deletions (96.4%) were shorter than 100 bp and excluded from further processing. Length distributions of the deletion callsets showed distinct peaks of 200 bp and 6 kb, corresponding to full-length copies of SINEs and LINE1s, respectively

(supplementary figs. 5 and 6, Supplementary Material online). After filtering, we retained 12,865 (Pindel) and 296,013 (BreakDancer) high-quality deletion calls that were between 100 bp and 10 kb long (supplementary tables 9 and 10, Supplementary Material online).

The majority (95%) of detected Pindel deletions were also identified by BreakDancer, suggesting a higher reliability at the expense of lower sensitivity in the program Pindel. The filtered data of both programs were merged into a nonredundant set of 295,434 deletion calls (supplementary table 11, Supplementary Material online). Of these, 270,689 (92%) matched TE annotations in the polar bear genome and hence were considered as Ref+ TE insertions. We detected 210,999 deletions that intersected SINE insertions in the polar bear genome. From 30,609 deletions matching LINE1 insertions, only a minor fraction (2.5%) was longer than 5 kb, the remaining copies were likely 5′-truncated (supplementary table 11, Supplementary Material online).

Phylogenetic networks generated from Ref+ and Ref− data sets, respectively, show that one type of detected insertions can only resolve one side of the phylogenetic tree (supplementary figs. 7 and 8, Supplementary Material online).

TE Insertion Rates in Ursine Bears

For both Ref+ and Ref− insertions, TeddyPi discovered on an average 10,000 and 20,000 TE insertions per genome, respectively (fig. 2*a*). The few TE insertions discovered in the two resequenced polar bears reflect the species' low genetic diversity and are expected because the reference genome is of a conspecific individual. Compared with LINE1 insertions, novel SINE insertions were ~6 times more frequent. TeddyPi identified 1.5 times more Ref+ than Ref− insertions in the bear genomes (fig. 2*a*). The highest number of TE insertions was found in the spectacled bear and the lowest number of TE insertions was identified in the two additional polar bear genomes (fig. 2*a*). For the other species, the numbers of identified TE insertion were homogeneous. As expected from their higher abundance, the genomic distance between SINE insertions was shorter than for LINEs (median distance: 10,010 and 73,240 bp, respectively) (fig. 2*b*). For the distance between SINEs, the upper bound was 330 kb. The upper bound of the LINE1 distances of >1 Mb indicates the presence of large genomic regions in which TeddyPi did not detect ursine-specific LINE1 insertions.

The rate of TE mobilization is known to differ between lineages (Hormozdiari et al. 2013). Among bears, LINE1-mediated retrotransposition of LINEs and SINEs is ubiquitous, but insertion rates (i.e., the number of TE insertion per generation) were substantially higher in brown and polar bear (fig. 2*c*). With 0.12 SINE insertions per generation, the insertion rate in the brown bear genome was the highest. TE insertions into coding or regulatory regions disrupt reading frames or inhibit transcription, however beneficial and potentially adaptive TE insertions are known (Cordaux and Batzer 2009; Casacuberta and Gonzalez 2013; Hof et al. 2016). In bears, 97% of TE insertions integrated into noncoding regions and only a few are located in exons or potential regulatory regions (supplementary fig. 9, Supplementary Material online).

In Vitro Validation of the TE Prediction Accuracy

Predicting TE insertions from high-throughput sequencing data is challenging and prone to artifacts. We extracted 151 loci to perform validation assays using PCR and Sanger sequencing to assess the accuracy of the in silico predictions (supplementary data 1, Supplementary Material online). All Sanger-sequenced loci, for which the size of the PCR amplicon suggested a TE insertion, were validated as a SINEC1_Ame insertion. Furthermore, the target site duplications and breakpoints were identical among different taxa, thus indicating a single, unique integration event (supplementary fig. 10 and note 1, Supplementary Material online). The validation experiment showed that 90% of the Ref− TE calls were accurate and both, false positive (FPR) and false negative rates (FPR) were low (table 1). The results indicate that the Ref− callers are more likely to miss a true TE insertion than to return an artifact. Loci were randomly selected for PCR validation from the whole data set or predefined presence/absence patterns for phylogenetic hypotheses (supplementary table 2, Supplementary Material online). Irrespectively, of whether the hypothesis matched the species tree or is in conflict with it, 93% of the predictions were experimentally confirmed to be accurate (table 1, supplementary data 1, Supplementary Material online).

In all 40 verified Ref+ TE insertion loci, an insertion was present in the polar bear genome, proving the reliability of our approach to select for Ref+ TE insertions. Prediction accuracy for Ref+ insertions in other species was 74% mainly attributed to a higher FPR than in Ref− insertions. A false positive Ref+ TE insertion call means that deletions were not recovered by SV callers, therefore Ref+ FPR should be considered as FNR.

For 111 loci, the PCR amplification yielded an unambiguous phylogenetic informative signal, that is, amplicon size differences with amplification success in all species. For 40 additional loci, one or more individual did not yield a PCR amplicon, and the locus was recorded as inconclusive. For all in vitro validated loci, we identified 17 loci with heterozygous SINE insertions (supplementary table 12, Supplementary Material online). In the brown bear, 17% of the amplified insertions were heterozygous. For the American black, Asiatic black, sun, sloth, and polar bear TE heterozygosity was 6% or less.

Interestingly, two SINE insertions (No. 40 and 122) were present in all ursine species except the polar bear. The flanks around the empty insertion site in the polar bear lack deletions and only the preintegration site was present compared with the other ursine bears. Other validated species-tree

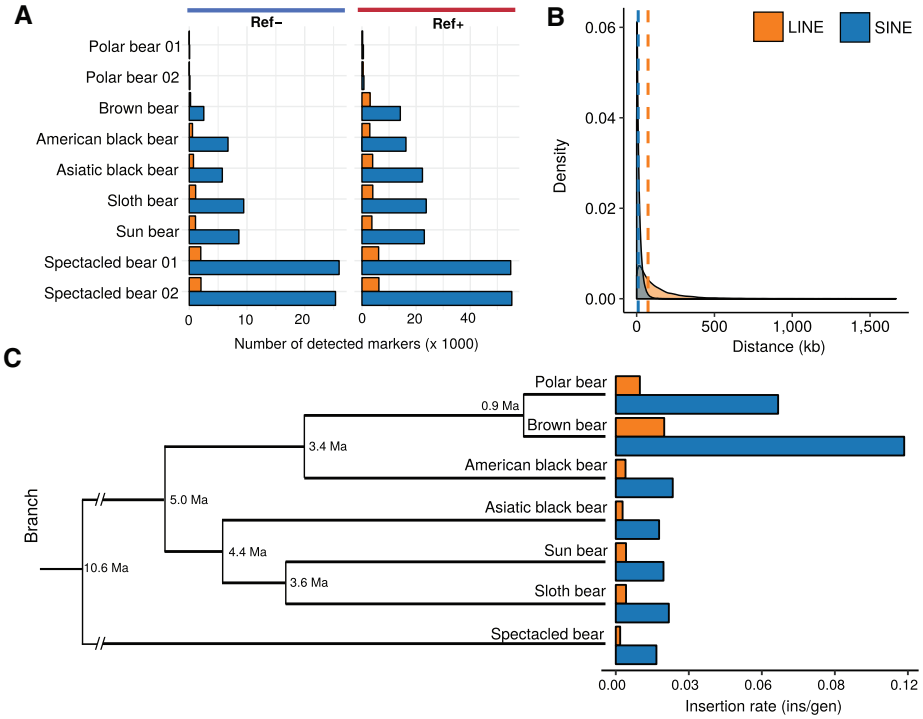


Fig. 2.—Detection results for TE insertions calls and inferred TE insertion rates. (*a*) Counts of Ref − (left) and Ref + (right) TE calls per analyzed sample shown for long interspersed element (LINE) insertions (orange) and short interspersed element (SINE) insertions (blue). (*b*) Distance distribution of all detected TE insertions among all bears. Vertical dashed lines indicate median distances. (*c*) TE insertion rates as insertions per generation (ins/gen) for all ursine species were estimated for the terminal branches in a chronogram scaled to divergence times from Kumar et al. (2017).

Table 1

Summary of In Vitro TE Validation Experiments for Ref− and Ref+ Insertion Loci

Type	Set	Informative Loci				All Loci			
		N	TP	FP	FN	N	TP	FP	FN
Ref−	All Ref−	80	0.90	0.04	0.06	111	0.87	0.07	0.06
	Hypothesis-driven	48	0.93	0.03	0.04	71	0.88	0.05	0.07
	Random	32	0.82	0.05	0.06	40	0.80	0.13	0.07
Ref+	All Ref+	31	0.74	0.23	0.04	40	0.70	0.26	0.03
	Pindel+BreakDancer	17	0.76	0.23	0.02	20	0.71	0.28	0.01
	Pindel	8	0.79	0.14	0.07	10	0.70	0.24	0.06
	BreakDancer	6	0.67	0.31	0.02	10	0.71	0.27	0.14

Note.—Results are shown for loci that were phylogenetically informative and all loci, that is, those lacking amplicons in more than one sample (All). The number of tested loci (N) and frequency of amplicon size differences that matched the computational prediction (true positives, TP), and false positively (FP) or false negatively (FN) predicted insertions are shown. For Ref− loci, random loci (Random), and loci predicted to support a specific phylogenetic hypothesis (Hypothesis-driven) were selected. For Ref+ markers, all loci were randomly selected.

incongruent TE insertions (supplementary fig. 11, Supplementary Material online) support alternative tree topologies reflecting the mitochondrial phylogeny or previously identified gene-flow signals from individual gene trees (Yu et al. 2007; Kutschera et al. 2014; Kumar et al. 2017). For example, seven validated TE insertions were synapomorphic for American and Asiatic black bear and nine insertions were shared by Asiatic black bear and sloth bear.

Reconstructing the Phylogeny of Bears

The Ref+ and Ref− TE insertions were merged into a common data set comprised of 150,513 SINE and LINE1 insertions. From these, 71,444 (47.5%) of the TEs were phylogenetically informative and 70,356 (46.7%) were species-specific. We found 8,713 TE insertions being shared by all seven bear species. However, the numbers of shared TE insertions differ when applying maximum parsimony that accounts for missing data (fig. 3).

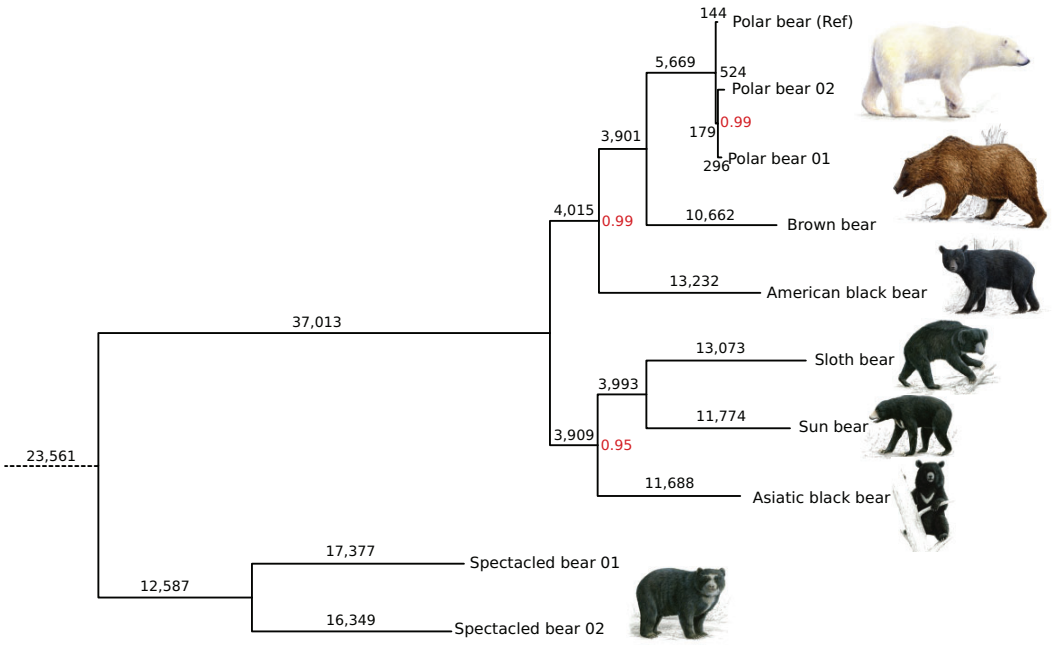


Fig. 3.—Dollo parsimony tree of bears reconstructed from 132,039 SINE insertions. Parsimony inferred branch lengths indicate the number of SINE insertions on that branch. Most nodes received bootstrap support of 100% (not indicated). Bootstrap support <100% is shown in red. The rescaled consistency index is 0.567, indicating conflict in the data set.

We identified seven times more insertions of SINEs than LINEs (132,093 and 18,420, respectively) across the bear genomes. The phylogenetic analysis focused on SINE insertions because these are shorter than the mean insert size of the sequencing libraries and thus robustly recovered by TE and SV calling. A Dollo parsimony analysis of 132,093 SINE insertions resulted in a phylogenetic tree with 100% bootstrap support for all nodes, except for the node separating the two polar bear individuals (fig. 3). The tree clearly groups spectacled bears that belong to the family Tremarctinae, outside the ursine bears. Within Ursinae, the tree has two clades that consist of the polar, brown, and American black bear and the Asiatic black, sun and sloth bear, respectively. Sun and sloth bear form a sister group to the Asiatic black bear. Despite, having 100% bootstrap support and branches that are generally supported by several thousand independent SINE insertions, a rescaled consistency index of 0.567 indicated phylogenetic incongruence among the data.

To explore phylogenetic conflict, a network analysis of the same data revealed a tree-like network. Similarly to the Dollo parsimony tree, the network clearly separated the Asiatic black, sloth, and sun bear from the other three ursine bears by a long edge, that represented 3,305 SINE insertions (fig. 4). Still, strong conflict among the Asiatic black, sun, and sloth bear was indicated by an intertwined web, that also included common splits with the polar or brown bear. Polar and brown bear were grouped by an edge that represents 3,597 SINE insertions, but polar bears also shared 2,240 insertions with the American black bear.

Phylogenetic conflict can be caused by hybridization or ancient polymorphisms that lead to allele sharing between nonsister group lineages and has been demonstrated for different ursine bears (Kutschera et al. 2014; Kumar et al. 2017). We stringently analyzed the phylogenetic conflict among Asiatic black, sun, and sloth bear using shared SINE insertions obtained from the presence/absence matrix without allowing for any missing data. The Asiatic black bear shares 278 SINE insertions with the sun bear and 265 SINE insertions with sloth bear. The monophyly of sun and sloth bear is supported by 168 SINE insertions. For these three taxa, statistical analyses using the KKSC test (Kuritzin et al. 2016) support the species-tree topology at high significance (bifurcation test, $P = 2.325e\text{-}10$) and reject hybridization between sun bear and the Asiatic black bear (hybridization test, $P = 0.6060$, supplementary table 13, Supplementary Material online). For the American and Asiatic black bear, 129 shared SINE insertions were recovered (fig. 5*b*), however the statistical significance of this result could not be assessed with existing methods.

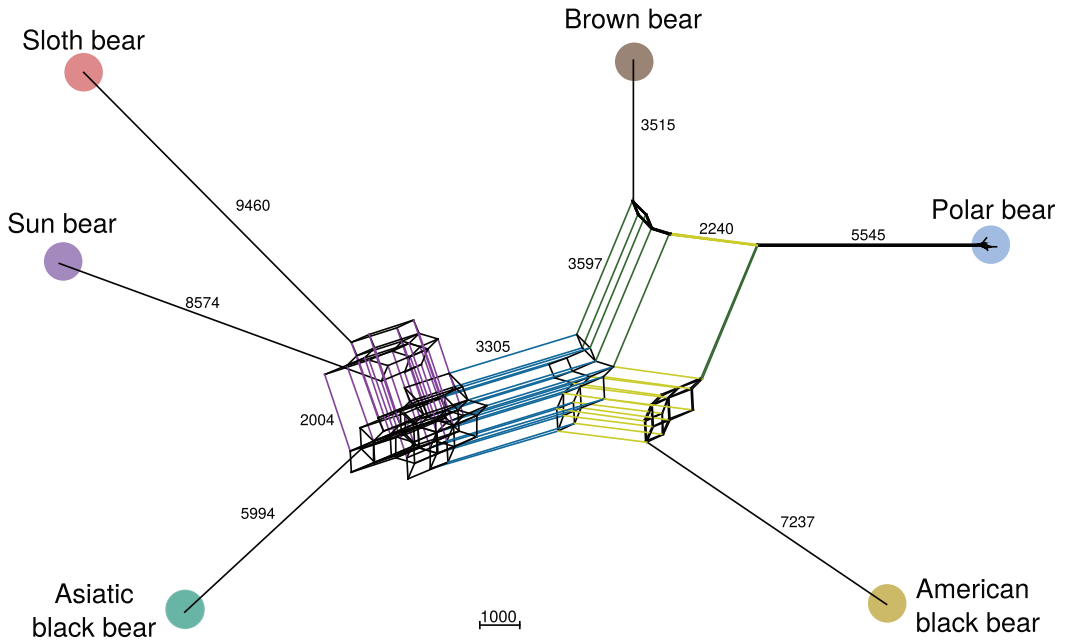


FIG. 4.—Median network from 132,093 SINE insertions. Parallel edges indicate shared splits between species. Major edges are colored, they separate the two major ursine clades (blue), or group together sun and sloth bear (purple), brown bear and polar bear (green) and American black and polar bear (yellow). Edge lengths indicate the number of shared SINE insertions as calculated by SplitsTree 4. For better readability the spectacled bear is not shown.

The monophyly of polar and brown bear is supported by 3,160 SINE insertions and the species-tree topology of polar, brown, and American black bear is significantly supported (tree test, $P = 1.04e\text{-}159$). The monophyly of all three species is supported by 2,178 SINE insertions (supplementary fig. 12, Supplementary Material online).

Different Extent of Phylogenetic Signal in the Flanking Regions

Alignments of genomic sequences flanking phylogenetically informative TE insertion sites were analyzed for their phylogenetic signal as well as for congruence with the phylogenetic signal from the adjacent TE insertion. Up to 65% of the individual ML trees calculated from the flanking sequences were identical with the presence/absence pattern of the TE insertion (fig. 6). To investigate the spatial congruence between the TE insertion and its flanks in more detail, we measured the number of substitutions that reconstructed the same phylogeny as the TE insertion in 1-kb nonoverlapping windows extending up to 10 kb from the insertion site (fig. 6). TE supporting substitutions were elevated in the direct vicinity of the TE insertion site and then tapered off with distance from the insertion site. The frequency of supporting substitutions is highest at TE insertion sites that are congruent with the ursine species tree and lower for those with a conflicting signal. For example, among 215 orthologous TE insertions shared by all Asiatic bears, the average frequency of TE-supporting substitutions increased from 0.01 to 0.04 within the first 5 kb from both sides of the insertion site (fig. 6). For species-tree incongruent TE insertion loci, the elevation of TE-supporting substitutions was less pronounced and the stretch of spatial congruence was shorter. Substitution frequencies for phylogenies that are different to the TE insertion signal were generally not elevated toward the insertion site (supplementary fig. 13, Supplementary Material online). In cases of only a minor difference in the phylogenetic signal between substitutions and TE, substitution frequencies were increased (supplementary note 2, Supplementary Material online).

Discussion

Analyzing whole genome sequence data for TE insertions makes it possible to study the landscape of genetic variation at unprecedented extent and detail. However, it faces methodological challenges. Here, we developed the TeddyPi pipeline that integrates different available TE callers and applies stringent filtering to overcome limitations of TE calling. It produces an automated output of presence/absence tables of TE insertions that can be immediately used for phylogenetic

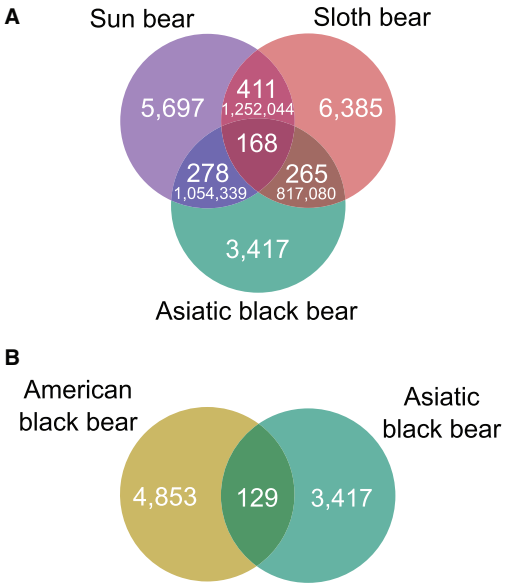


Fig. 5.—Venn Diagrams depicting phylogenetic conflict among Asiatic black, sun, and sloth bear (*a*) and American black and Asiatic black bear (*b*). The amount of shared SINE insertions under Dollo parsimony are shown. The numbers in smaller font give the number of shared genome-wide nucleotide substitutions calculated using the *D*-statistics (Kumar et al. 2017).

analyses. The pipeline follows a "quality over quantity" approach to select highly reliable TE insertion loci. Recent phylogenomic studies suggest that genomes are often a mosaic of different genealogies caused by evolutionary processes such as introgressive hybridization or ILS (Mallet et al. 2016). To study such complex signals, sufficient character sampling is necessary. This can only be achieved by nucleotide-based genome analyses, or genome-wide and unbiased discovery of TE insertions (Kuritzin et al. 2016; Dodt WG, Gallus S, Matthew PJ, Nilsson MA, Unpublished data). TE insertion data provide an independent and robust molecular marker system to build phylogenies that are not based on sequence analysis (Shedlock et al. 2004).

SINE Insertions Recapitulate the Evolutionary History of Bears

Extensive phylogenetic discordance across loci has previously challenged the resolution of the bear phylogeny (Yu et al. 2007; Kutschera et al. 2014; Kumar et al. 2017). The TeddyPi pipeline extracted 132,093 SINE insertions from low-coverage data to build a reliable data set of phylogenetically informative TE markers to study the evolutionary history of bears. We reconstructed a well-supported phylogenetic species tree despite incongruent phylogenetic signals (figs. 3 and 4). The three Asian bears form a clade that is consistent with coalescent analyses of genome sequence data (Kumar et al. 2017). However, this contrasts with previous studies, that placed the Asiatic black bear as sister group to the polar, brown, and American black bear clade or as sister group to the American black bear, respectively (Yu et al. 2007; Krause et al. 2008; Pagès et al. 2008). Despite significant bootstrap support for each node of the parsimony TE tree, the tree had a low-consistency index, indicating that many TE insertions conflict with the inferred phylogeny. Phylogenetic networks can depict such conflicting signals better than trees that force the data to a bifurcating model of evolution (Bapteste et al. 2013). The network analyses reveal that phylogenetic conflict among bears occurs mostly in the two main clades of the ursine subfamily (fig. 4). In particular, the Asiatic black, sun, and sloth bear that currently inhabit South-East Asia form a complex network. We explored this conflict further and found that the Asiatic black bear share almost identical numbers of orthologous SINE insertions with sun and sloth bear, respectively, thereby indicating ILS as the origin of the conflict (fig. 5 and supplementary table 13, Supplementary Material online). Despite reconstructing the same species tree, our detailed analyses contrast nucleotide-based analyses of millions of sites, that inferred ancestral hybridization as the main driver of phylogenetic conflict among these species (Kumar et al. 2017). To what extent hybridization occurred between bears and what caused the conflicting signal of single nucleotide substitutions and TE insertions remains to be further explored.

In previous mtDNA-based analyses, the Asiatic and American black bear have been placed as sister species (Yu et al. 2007; Krause et al. 2008). This is not supported by the majority of identified TE insertions. However, 129 SINE insertions are shared by American and Asiatic black bear (fig. 5). Therefore, the close relationship of the two black bears based on mtDNA analyses is likely a result of an ancient mitochondrial capture event and additional introgression of nuclear DNA carrying these TE insertions (Kutschera et al. 2014). An alternative scenario explaining the discordance between mtDNA and nuclear DNA phylogenies of American and Asiatic black bear involves nuclear swamping of the American black bear genome by brown bear alleles. In this scenario, the mitochondrial phylogeny reflects the true speciation history but was eventually obfuscated by introgression of brown bear DNA into the American black bear genome. This would produce a similar phylogenetic signal and artificially place the American black bear on the lineage leading to brown and polar bear (Kutschera et al. 2014). However, our network analysis and 99 SINE insertions shared by brown bear and American black bear give very little support for this hypothesis, suggesting that ancient hybridization between the two black bear species had a more pronounced effect on their genomes than nuclear swamping by brown bear DNA (fig. 3 and supplementary fig. 12, Supplementary Material online). If

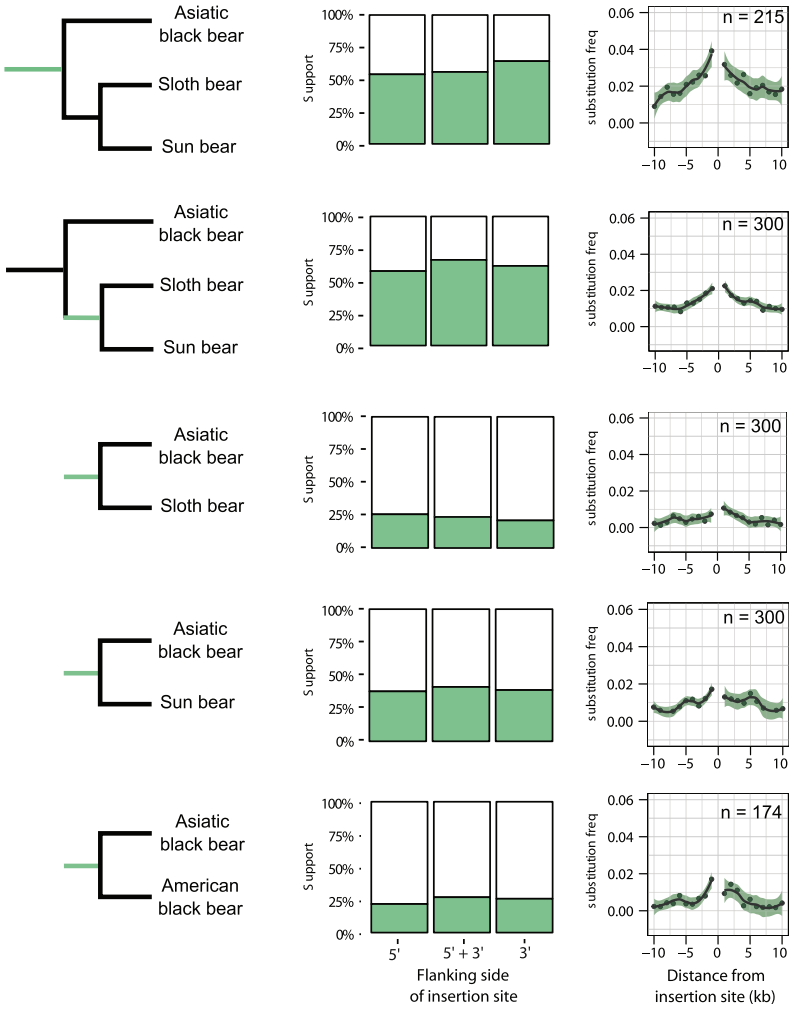


Fig. 6.—Analysis of flanking sequences of TE insertions present in different groups of taxa. Left panel: Green branches in the phylogenetic tree indicate when the TEs integrated. Middle panel: Bar plots showing the frequency of ML-trees calculated from 10-kb flanking sequence on the 5′, 3′ end or a concatenation of both. Right panel: Frequency of substitutions that support the TE insertion signal in 1-kb windows around the insertion site. Frequencies are normalized by the number of segregating sites.

the phylogenetic conflict during the initial radiation of Ursinae was caused by ILS, approximately equal number of TE insertions supporting different evolutionary scenarios were expected. This is not evident from our analyses.

Differences in retrotransposition activity or demographic history can cause varying rates of TE insertions between lineages (Hormozdiari et al. 2013). Our insertion rate estimates were 0.022 SINE and 0.004 LINE1 insertions per genome per generation, which is half of the rate for humans (0.035 Alus and 0.008 LINE1s) (fig. 2c; Sudmant et al. 2015). Fixation of neutral or slightly deleterious TE insertions depends on genetic drift, which is stronger in small effective population sizes or on purifying selection, which is stronger in large populations (Charlesworth 2009; Gonzalez and Petrov 2012). The substantially higher insertion rates of TEs and the high heterozygosity rate in brown bear can thus be explained by the large population size that brown bears have maintained over long timespans (Miller et al. 2012). Polar bears exhibit a low heterozygosity of TE insertion, reflecting the species low genetic diversity as consequence of population

bottlenecks (Hailer et al. 2012; Miller et al. 2012). In polar and brown bear, TE insertion rates are higher than in the other bears. The high TE insertion rate in polar and brown bears can also be explained by a retrotranspositional burst caused by hybridization (O'Neill et al. 1998; Dion-Côté et al. 2014). Bears are known to hybridize, and hybrids between polar and brown bears have been observed (Galbreath et al. 2008; Kelly et al. 2010). Additionally, a hybrid origin of polar bears has been proposed (Lan et al. 2016). Thus, consequent genetic introgression potentially leads to a burst of TE insertions in the species into which hybrids backcross and thus may explain the high TE insertion rate in brown and polar bears. This indicates, that there is no general genomic mechanism to suppress genomic insertions as was suggested by the absence of mitochondrial pseudogene insertions (Lammers et al. 2017).

The accompanying sequence-based analyses of the same data set enabled to examine the correlation of nucleotide substitutions and TEs for conflicting phylogenies (Kumar et al. 2017). Expectedly, TE insertions were several magnitudes less frequent than nucleotide substitutions. Yet, both analyses yielded the same phylogeny but differed in their interpretation of phylogenetic conflict (fig. 5). Huff et al. (2010) described that TE(s) have on an average older genealogies due to the rarity of TE insertion compared to the nucleotide mutation rate. Thus, TE loci have deeper coalescence times and a higher probability for ILS (Maddison 1997). Also, introgression of alleles carrying TE insertions might be less frequent because they can be deleterious due to genetic incomparability, that is, Dobzhansky–Muller incompatibilities (Dobzhansky 1941; Muller 1942). This highlights the need for nucleotide-based analyses in addition to genome wide analyses of TE insertions.

Quality over Quantity Approach for Phylogenetic Inference of TEs

Previous phylogenetic TE analyses relied on the availability of reference genomes which were often restricted to one species per order or family. For bears, draft genome assemblies of polar bear and giant panda are available (Li et al. 2010; Liu et al. 2014). Traditional in vitro approaches would have identified orthologous loci in both genomes, with one carrying a TE insertion that is experimentally tested for presence or absence in the other bears using PCR (Shedlock et al. 2004). Although the availability of two reference genomes is beneficial, unbiased identification of variable, that is, phylogenetically informative TEs across the complete taxon-sampling is not possible using this approach. Adding genomes from the entire ursine subfamily makes it possible to discover TE insertions free from sampling artifacts and to precisely extract phylogenetically informative markers. However, the nested position of the polar bear reference genome inside the species tree, the use of low-coverage genome data and misassembled regions in the reference genome were challenging for TE calling and required methodological refinements to increase prediction quality of TE insertions. These challenges were rarely discussed in other studies but are central when aiming for a large-scale identification of TE insertions from paired-end mapping data without introducing a sampling bias.

If the reference genome is nested inside the ingroup, as in the case of the polar bear inside Ursinae, a two-sided approach using Ref+ and Ref− insertions is necessary to yield support for all internodes in the resulting phylogenetic tree or network (supplementary fig. 2, Supplementary Material online). The polar bear genome sequence has a higher contiguity than that of the giant panda, a better assembly of repeats due to longer sequencing reads and it benefits from the low heterozygosity in polar bear. Compared with the polar bear reference genome, the giant panda genome is less suited to be used as a reference for mapping because of its high evolutionary distance to the other bear species, which diverged from the giant panda some 20 Ma. To solve this problem and to make TeddyPi more ubiquitously applicable, SV callers were integrated in the pipeline to deduce Ref+ insertions from deletions calls (Nellåker et al. 2012). Only few TE callers are specifically developed to detect Ref+ insertions. To our knowledge, only T-lex and T-lex2 (Fiston-Lavier et al. 2011, 2015) perform Ref+ insertion detection, but they are not compatible with the TeddyPi pipeline due to different file format requirements. Other programs, such as RetroSeq, Mobster and Jitterbug exclusively detect Ref− TE insertions (Keane et al. 2013; Thung et al. 2014; Hénaff et al. 2015). Depending on the mapping-signature utilized for SV-calling (split-reads, read-pairs, depth of coverage) detection results differed markedly between programs as exemplified by our results from Pindel and Breakdancer (supplementary tables 9 and 10, Supplementary Material online) and by results from other studies (Ewing 2015). Inconsistencies between different programs will affect the phylogenetic inference, which relies on precise presence/absence patterns of orthologous loci and make it necessary to integrate different SV callers as implemented in TeddyPi. Despite the general concordance of TE calls from Mobster and RetroSeq, only overlapping calls were used to increase the reliability of the calls. For TE calling, integration of multiple callers is recognized as an appropriate strategy to enhance the consistency of TE predictions (Lin et al. 2015; Nelson et al. 2017), and this functionality is implemented in TeddyPi for both, Ref+ and Ref− insertions. A true positive rate (TPR) of 93% for TE calls from the TeddyPi pipeline (table 1) is higher than the estimated sensitivity of RetroSeq for 10× whole genome sequencing data (Keane et al. 2013). The reliability of TeddyPi is equally good as estimates from Mobster analyses of high-quality human data (Thung et al. 2014). The false positive rate for Ref− TE calls is low (4%), but considerably higher for Ref+ insertions (23%). Thus, when possible, the use of a suitable outgroup genome to analyze only Ref− insertions for phylogenetic reconstruction is recommended.

Detecting TE insertions and SVs in resequenced whole genome data often have breakpoint inaccuracies within a margin of up to 50 bp (Ewing 2015). It is therefore not possible to distinguish between near or near-exact deletion or insertions. This can affect detecting ortholog events or analyzing genetic effects by intersection with coding sequences or regulatory regions (supplementary fig. 7, Supplementary Material online). Given the short length of regulatory sequences an overestimation of disrupting TE insertions can not be excluded. However, breakpoint inaccuracies are unlikely to have affected the detection of orthologous TE insertions because long near-exact indels occur at a very low level (van de Lagemaat et al. 2005). Therefore, they would have contributed only marginally to the observed phylogenetic conflict among bears.

Missing data and unplaced scaffolds are common in most genome assemblies, because of current technological limitations to sequence and assemble repetitive DNA. Thus, in genome sequences, sequence gaps are mostly caused by repetitive regions, such as TEs and satellite DNA. Long read sequencing technologies, such as PacBio or Nanopore, are expected to alleviate this problem considerably. The 2.3 Gb polar bear genome sequence was based on short read technology and based on an estimated genome size of 2.7 Gb for extant bears lacks 400 Mb of genomic information (Vinogradov 1998; Krishan et al. 2005; Liu et al. 2014). Another artifact from repetitive DNA in genome sequences are unassembled regions in the scaffolds (N-regions). TeddyPi utilized 38 Mb of N-regions in the polar bear genome as a proxy for poorly assembled regions, and all TE calls in their vicinity were excluded from the analyses. The removal of N-regions greatly increased the success rates in the experimental validation and show that this is a necessary step in TE calling, that previously has not been implemented. Another indicator of assembly quality and of the ability to confidently predict TEs is the mappability (or uniqueness) of short-reads to the reference genome. Mappability can be assessed by deviations of local coverage depth from the mean coverage. To account for poorly mapped regions, TE calls in regions of exceptionally low- and high coverage were coded as missing data. Another challenge to TE and SV calling comes from the random integration of TEs in the genome. Occasionally, TEs can randomly integrate into older TE sequences. If both TEs are of the same type, sequence reads will be ambiguously mapped to either the young or old TE. This increases the risk for false positive calls during TE calling. Therefore, TE calls located within annotated TEs of the same type were removed in the TeddyPi pipeline to increase the reliability of our phylogenetic markers.

Unlike for the human genome, a generally accepted standard or database of TE insertions does not exist for nonmodel organisms to compare our results to. Thus, detection sensitivity can only be estimated by experimental approaches. The validation experiments show that compared with standard TE callers, the rigorous approach of the TeddyPi pipeline substantially improves TE detection from nonmodel organism genomes that lack highly curated and well-annotated genome assemblies. For the polar bear genome sequence, every experimentally verified locus was confirmed for the presence of SINEC1_Ame, corroborating the assembly and RepeatMasker annotation for these loci. The presence of TSDs in all analyzed loci further strengthens the TeddyPi approach in identifying true, orthologous TE insertion events.

TE Insertions, Flanking Sequences, and Recombination Blocks in Ursine Bears

TE insertions share an evolutionary history with nucleotide substitutions occurring in their immediate genomic vicinity (Daly et al. 2001). If the TE insertion is neutral, the extent of linkage, that is, the size of a recombination block that carries the TE depends on the recombination rate and the demographic history of the genomic region (Ellegren and Galtier 2016). In great apes, phylogenetic congruence between the TE insertion and its flanking sequence was used to prove hemiplasy of the TE insertion (Hormozdiari et al. 2013), however nucleotide-homoplasy and uncertainties in tree-reconstruction of the specific regions can mislead such an analysis, especially for longer timescales (Suh et al. 2015). Ursine bears radiated ~5 Ma, which left little time for flanking sequences to be saturated, allowing for nucleotide level comparisons. In bears, TE insertions and their flanking sequences share the same phylogenetic signal, but the extent of spatial congruence (i.e., linkage) is limited to a few kb and differs depending on the phylogenetic signal of the TE (fig. 6 and supplementary fig. 12, Supplementary Material online). The size of the recombination block, as evident from the extent of spatial congruence (fig. 6), gives a relative estimate of the time since the TE insertions. A lesser extent of spatial congruence around the species-tree incongruent TE insertions can be explained by an earlier TE integration and subsequent breakdown of the recombination blocks. TE insertions shared exclusively by American and Asiatic black bear have a narrow extent of spatial congruent substitutions, and thus are older than species-tree congruent TE insertions. If a locus originates from more recent introgression a wider extent of spatial congruence carrying the same phylogenetic signal is expected. The flanks of the orthologous TE insertions in the Asiatic bears share the same phylogenetic signal, and therefore show no homoplasy and suggest that ILS has contributed to the phylogenetic incongruence among these loci. For the Asiatic bears, we propose that ILS is the primary driver of phylogenetic incongruence causing high amounts of pairwise similarities (fig. 5*a*; Kutschera et al. 2014) and additionally, hybridization between Asiatic black and sun bear led to an excess of shared alleles between these species (fig. 5*a*; Kumar et al. 2017). Under the assumption that the current species tree of bears (fig. 3) reflects the speciation history, introgressive hybridization involving the American black bear must have

occurred. However, in agreement with coalescent-based analyses (Kutschera et al. 2014), analyses of TE insertion patterns and their flanking regions (figs. 5 and 6) indicate that the local geneologies are not yet sorted, thereby confounding introgression analyses. Although our sequence analyses of the TE flanking regions were restricted to one taxonomic group, it is evident that analyses of deeper divergences in any taxa will have shorter recombination blocks and thus fewer phylogenetic signatures. Thus, screening for flanking substitutions surrounding old TE insertions is likely to be uninformative due to the limited spatial congruence coupled with nucleotide saturation.

Conclusion

Twenty years after the successful introduction of TE insertions as phylogenetic markers, it is now possible to not only use a few but thousands of informative loci across the genome to reconstruct phylogenies of complete taxonomic groups. The TeddyPi pipeline allowed us to detect TE insertion in silico from nine bear genomes. Over 130,000 SINE insertions show that TE insertions are a major driver of genomic variation among ursine bears and reconstructed their phylogeny with virtually homoplasy-free evolutionary information. The TE phylogeny of bears confirm the presence of two distinct clades among Ursinae and significantly shows that Asiatic black, sun, and sloth bear form a monophyletic clade, despite a high degree of ILS. The conceptual framework of the integrated and stringent approach in TeddyPi allows an unbiased analysis of ancestry-informative TEs as a routine procedure in comparative genomic studies. Deciphering recent and complex speciation processes using TE insertions as well as nucleotide substitutions is subject to further analyses and important for our understanding of phylogenetics and speciation (Mallet et al. 2016).

Data Availability

The final TE data set, and primers for validation experiments are included as Supplementary Material online. TeddyPi is available at https://github.com/mobilegenome/teddypi, last accessed September 2017.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

The authors thank Dije Tjwan Thung and Jayne Hehir-Kwa, The Radboud University Medical Center, for providing an unpublished version of Mobster, Thomas Keane, Wellcome Trust Sanger Institute, for advice in using RetroSeq, and Markus Pfenninger for helpful discussions. The authors are thankful to Kathinka Schulze and Clara Heumann-Kieser for performing validation experiments, Alison Eyres for English proof-reading, and five anonymous reviewers for helpful comments on earlier versions of this manuscript. Jón Baldur Hlíðberg (www.fauna.is) painted the bears in figure 3. The publication of this article was funded by the Open Access Fund of the Leibniz Association.

Author Contributions

F.L., M.A.N., and A.J. conceived and designed the study. F.L. developed TeddyPi and performed the computational analyses. S.G. and M.A.N. coordinated and performed experimental validation experiments. F.L. and M.A.N. wrote the manuscript with input from all co-authors. All authors read and approved the final manuscript.

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215(3):403–410.
Bapteste E, et al. 2013. Networks: expanding evolutionary thinking. Trends Genet. 29(8):439–441.
Bidon T, et al. 2015. Genome-wide search identifies 1.9 Mb from the Polar Bear Y chromosome for evolutionary analyses. Genome Biol Evol. 7(7):2010–2022.
Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120.
Casacuberta E, Gonzalez J. 2013. The impact of transposable elements in environmental adaptation. Mol Ecol. 22(6):1503–1517.
Charlesworth B. 2009. Effective population size and patterns of molecular evolution and variation. Nat Rev Genet. 10(3):195–205.
Chen K, et al. 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods 6(9):677–681.
Churakov G, et al. 2009. Mosaic retroposon insertion patterns in placental mammals. Genome Res. 19(5):868–875.
Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. Nat Rev Genet. 10(10):691–703.
Cronin MA, Amstrup SC, Talbot SL, Sage GK, Amstrup KS. 2009. Genetic variation, relatedness, and effective population size of polar bears (*Ursus maritimus*) in the southern Beaufort Sea, Alaska. J Hered. 100(6):681–690.
Dale RK, Pedersen BS, Quinlan AR. 2011. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. Bioinformatics 27(24):3423–3424.
Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. 2001. High-resolution haplotype structure in the human genome. Nat Genet. 29(2):229–232.
Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet. 6(5):361–375.
Dion-Côté A-M, Renaut S, Normandeau E, Bernatchez L. 2014. RNA-seq reveals transcriptomic shock involving transposable elements reactivation in hybrids of young lake whitefish species. Mol Biol Evol. 31(5):1188–1199.
Dobzhansky T. 1941. Genetics and the origin of species. 2nd ed West Sussex: Columbia University Press.
Doucet AJ, Droc G, Siol O, Audoux J, Gilbert N. 2015. U6 snRNA pseudogenes: markers of retrotransposition dynamics in mammals. Mol Biol Evol. 32(7):1815–1832.
Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for ancient admixture between closely related populations. Mol Biol Evol. 28(8):2239–2252.

Ellegren H, Galtier N. 2016. Determinants of genetic diversity. Nat Rev Genet. 17(7):422–433.
Ewing AD. 2015. Transposable element detection from whole genome sequence data. Mob DNA 6:24.
Fiston-Lavier A-S, Barron MG, Petrov DA, Gonzalez J. 2015. T-lex2: genotyping, frequency estimation and re-annotation of transposable elements using single or pooled next-generation sequencing data. Nucleic Acids Res. 43(4):e22.
Fiston-Lavier A-S, Carrigan M, Petrov DA, González J. 2011. T-lex: a program for fast and accurate assessment of transposable element presence using next-generation sequencing data. Nucleic Acids Res. 39(6):e36.
Galbreath GJ, Hunt M, Clements T, Waits LP. 2008. An apparent hybrid wild bear from Cambodia. Ursus 19(1):85–86.
Gonzalez J, Petrov DA. 2012. Evolution of genome content: population dynamics of transposable elements in flies and humans. In: Anisimova M, editor. Evolutionary genomics: statistical and computationla methods. Vol. 855. New York: Springer-Humana, p. 361–383.
Hailer F, et al. 2012. Nuclear genomic sequences reveal that polar bears are an old and distinct bear lineage. Science 336(6079):344–347.
Hallström BM, Janke A. 2010. Mammalian evolution may not be strictly bifurcating. Mol Biol Evol. 27(12):2804–2816.
Hénaff E, Zapata L, Casacuberta JM, Ossowski S. 2015. Jitterbug: somatic and germline transposon insertion detection at single-nucleotide resolution. BMC Genomics 16:768.
Hof AEV, et al. 2016. The industrial melanism mutation in British peppered moths is a transposable element. Nature 534(7605):102–105.
Hormozdiari F, et al. 2013. Rates and patterns of great ape retrotransposition. Proc Natl Acad Sci. 110(33):13457–13462.
Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Mol Biol Evol. 33(6):1635–1638.
Huff CD, Xing J, Rogers AR, Witherspoon D, Jorde LB. 2010. Mobile elements reveal small population size in the ancient ancestors of *Homo sapiens*. Proc Natl Acad Sci U S A. 107(5):2147–2152.
Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 23(2):254–267.
Jurka J, et al. 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 110(1–4):462–467.
Keane TM, Wong K, Adams DJ. 2013. RetroSeq: transposable element discovery from next-generation sequencing data. Bioinformatics 29(3):389–390.
Kelly BP, Whiteley A, Tallmon D. 2010. The Arctic melting pot. Nature 468(7326):891.
Krause J, et al. 2008. Mitochondrial genomes reveal an explosive radiation of extinct and extant bears near the Miocene-Pliocene boundary. BMC Evol Biol. 8:220.
Kriegs JO, et al. 2006. Retroposed elements as archives for the evolutionary history of placental mammals. PLoS Biol. 4(4):537–544.
Krishan A, et al. 2005. DNA index, genome size, and electronic nuclear volume of vertebrates from the Miami Metro Zoo. Cytometry A 65A(1):26–34.
Kumar V, et al. 2017. The evolutionary history of bears is characterized by gene flow across species. Sci Rep. 7:46487.
Kuramoto T, Nishihara H, Watanabe M, Okada N. 2015. Determining the position of storks on the phylogenetic tree of waterbirds by retroposon-insertion analysis. Genome Biol Evol. 7(12):3180.
Kuritzin A, Kischka T, Schmitz J, Churakov G. 2016. Incomplete lineage sorting and hybridization statistics for large-scale retroposon insertion data. PLoS Comput Biol. 12(3):e1004812.
Kutschera VE, et al. 2014. Bears in a forest of gene trees: phylogenetic inference is complicated by incomplete lineage sorting and gene flow. Mol Biol Evol. 31(8):2004–2017.
Lammers F, Janke A, Rücklé C, Zizka V, Nilsson MA. 2017. Screening for the ancient polar bear mitochondrial genome reveals low integration of mitochondrial pseudogenes (numts) in bears. Mitochondrial DNA B 2:251–254.
Lan T, et al. 2016. Genome-wide evidence for a hybrid origin of modern polar bears. BioRxiv. doi.org/10.1101/047498.
Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26(5):589–595.
Li R, et al. 2010. The sequence and de novo assembly of the giant panda genome. Nature 463(7279):311–317.
Lin K, Smit S, Bonnema G, Sanchez-Perez G, de Ridder D. 2015. Making the difference: integrating structural variation detection tools. Brief Bioinform. 16(5):852–864.
Lindblad-Toh K, et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438(7069):803–819.
Liu S, et al. 2014. Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. Cell 157(4):785–794.
Maddison WP. 1997. Gene trees in species trees. Syst Biol. 46(3):523–536.
Mallet J, Besansky N, Hahn MW. 2016. How reticulated are species? BioEssays 38(2):140–149.
Medvedev P, Stanciu M, Brudno M. 2009. Computational methods for discovering structural variation with next-generation sequencing. Nat Methods 6:13–20.
Miller W, et al. 2012. Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. Proc Natl Acad Sci. 109:E2382–E2390.
Muller HJ. 1942. Isolating mechanisms, evolution and temperature. Biol Symp. 6:71–125.
Nellåker C, et al. 2012. The genomic landscape shaped by selection on transposable elements across 18 mouse strains. Genome Biol. 13(6):R45.
Nelson MG, Linheiro RS, Bergman CM. 2017. McClintock: an integrated pipeline for detecting transposable element insertions in whole genome shotgun sequencing data. G3 Genes Genomes Genet. 7:2763–2778.
Nikaido M, Rooney AP, Okada N. 1999. Phylogenetic relationships among cetartiodactyls based on insertions of short and long interpersed elements: hippopotamuses are the closest extant relatives of whales. Proc Natl Acad Sci. 96(18):10261–10266.
Nishihara H, Maruyama S, Okada N. 2009. Retroposon analysis and recent geological data suggest near-simultaneous divergence of the three superorders of mammals. Proc Natl Acad Sci. 106(13):5235–5240.
O'Neill RJ, O'Neill MJ, Graves JA. 1998. Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. Nature 393:68–72.
Onorato DP, Hellgren EC, Van Den Bussche RA, Doan Crider DL. 2004. Phylogeographic patterns within a metapopulation of black bears (*Ursus americanus*) in the American southwest. J Mammal. 85(1):140–147.
Pagès M, et al. 2008. Combined analysis of fourteen nuclear genes refines the Ursidae phylogeny. Mol Phylogenet Evol. 47(1):73–83.
Platt RN, et al. 2015. Targeted capture of phylogenetically informative ves SINE insertions in genus Myotis. Genome Biol Evol. 7(6):1664–1675.
Pontius JU, et al. 2007. Initial sequence and comparative analysis of the cat genome. Genome Res. 17(11):1675–1689.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6):841–842.
Ray DA, Xing J, Salem A-H, Batzer MA. 2006. SINEs of a nearly perfect character. Syst Biol. 55(6):928–935.
Shedlock AM, Takahashi K, Okada N. 2004. SINEs of speciation: tracking lineages with retroposons. Trends Ecol Evol. 19(10):545–553.
Shimamura M, et al. 1997. Molecular evidence from retroposons that whales form a clade within even-toed ungulates. Nature 388(6643):666–670.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9):1312–1313.

Sudmant PH, et al. 2015. An integrated map of structural variation in 2,504 human genomes. Nature 526(7571):75–81.
Suh A, Kriegs JO, Donnellan S, Brosius J, Schmitz J. 2012. A universal method for the study of CR1 retroposons in nonmodel bird genomes. Mol Biol Evol. 29(10):2899–2903.
Suh A, Smeds L, Ellegren H. 2015. The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. PLoS Biol. 13(8):e1002224.
Sukumaran J, Holder MT. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26(12):1569–1571.
Swofford D. 2002. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sunderland, Massachusetts: Sinauer Associates.
Tallmon DA, Bellemain E, Taberlet P, Swenson JE. 2004. Genetic monitoring of Scandinavian brown bear effective population size and immigration.DeWoody, editor. J Wildl Manage. 68:960–965.
Thung DT, et al. 2014. Mobster: accurate detection of mobile element insertions in next generation sequencing data. Genome Biol. 15(10):488.
Untergasser A, et al. 2012. Primer3—new capabilities and interfaces. Nucleic Acids Res. 40(15):e115.
van de Lagemaat LN, Gagnier L, Medstrand P, Mager DL. 2005. Genomic deletions and precise removal of transposable elements mediated by short identical DNA segments in primates. Genome Res. 15(9):1243–1249.
Vinogradov AE. 1998. Genome size and GC-percent in vertebrates as determined by flow cytometry: the triangular relationship. Cytometry 31(2):100–109.
Walters-Conte KB, Johnson DLE, Allard MW, Pecon-Slattery J. 2011. Carnivore-specific SINEs (Can-SINEs): distribution, evolution, and genomic impact. J Hered. 102(Suppl 1):S2–S10.
Wong K, Keane TM, Stalker J, Adams DJ. 2010. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. Genome Biol. 11(12):R128.
Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25(21):2865–2871.
Yu L, Li Y-W, Ryder OA, Zhang Y-P. 2007. Analysis of complete mitochondrial genome sequences increases phylogenetic resolution of bears (Ursidae), a mammalian family that experienced rapid speciation. BMC Evol Biol. 7:198.

Associate editor: Ellen J. Pritham

Supplementary Information Document

Phylogenetic conflict in bears identified by automated genome-wide discovery of transposable element insertions

Fritjof Lammers[1,2], Susanne Gallus[1], Axel Janke[1,2], Maria A Nilsson[1§]

[1]Senckenberg Biodiversity and Climate Research Centre, Senckenberg Gesellschaft für Naturforschung, Senckenberganlage 25, 60325 Frankfurt am Main, Germany.
[2]Goethe University Frankfurt, Institute for Ecology, Evolution & Diversity, Biologicum, Max-von-Laue-Str.13, 60439 Frankfurt am Main, Germany.

Supplementary Figures

Supplementary Figure 1. Flowchart of the TeddyPi pipeline.

Supplementary Figure 2. Considerations for the choice of TE callers.

Supplementary Figure 3. The principle of deletion calling.

Supplementary Figure 4. Repeat landscapes from the genome assemblies of polar bear (A) and giant panda (B).

Supplementary Figure 5. Length distribution for deletions called by Pindel (dotted lines) and Breakdancer (solid lines) for each analyzed genome.

Supplementary Figure 6. Length distribution for deletions called by Pindel (dotted lines) and Breakdancer (solid lines) for each analyzed genome.

Supplementary Figure 7. Phylogenetic networks reconstructed from SINE insertions shown separately for Ref- insertions (A) and Ref+ insertions (B).

Supplementary Figure 8. Phylogenetic networks reconstructed from LINE1 insertions shown separately for Ref- insertions (A) and Ref+ insertions (B).

Supplementary Figure 9. Insertion frequency of TEs (SINEs and LINEs) in different genomic contexts in the polar bear genome.

Supplementary Figure 10. Alignment of marker 104.

Supplementary Figure 11. Phylogenetic signal from TE markers that are species-tree incongruent based on validation experiments.

Supplementary Figure 12. Venn Diagram showing conflict among Polar bear, brown bear and American black bear on basis of inferred SINE insertions using Dollo Parsimony.

Supplementary Figure 13. Phylogenetic signals in the genomic sequences flanking the TEs.

Supplementary Tables

Supplementary Table 1. List of genomes analyzed in this study.

Supplementary Table 2. Selected phylogenetic hypotheses subject to validation experiments.

Supplementary Table 3. Repetitive elements in the polar bear genome sequence.

Supplementary Table 4. Prediction counts from RetroSeq SINE calls for raw calls and each filtering step.

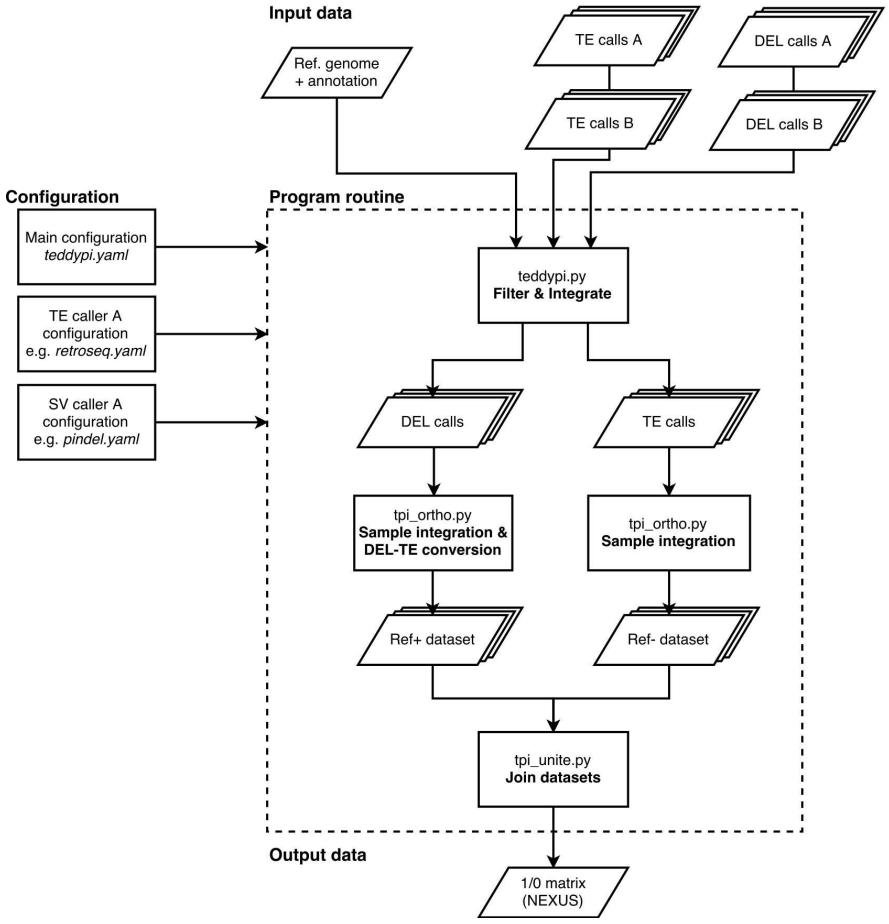
Supplementary Table 5. Predictions counts from Mobster for raw calls and each filtering step for SINEs.

Supplementary Table 6. Prediction counts from RetroSeq LINE1 calls for raw calls and each filtering step.

Supplementary Table 7. Predictions counts from Mobster for LINE1 calls and each filtering step.

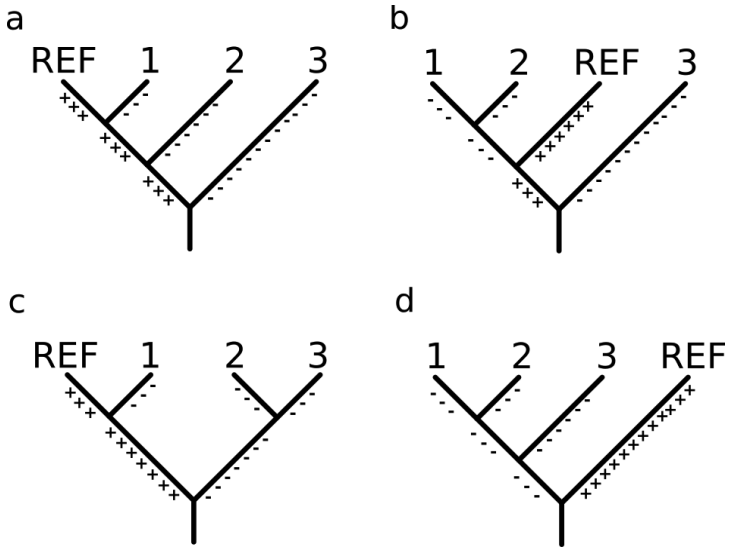
Supplementary Table 8. Summary of non-reference TE insertion counts in Ursinae for SINEs and LINEs with values from RetroSeq and Mobster and their overlap.

.

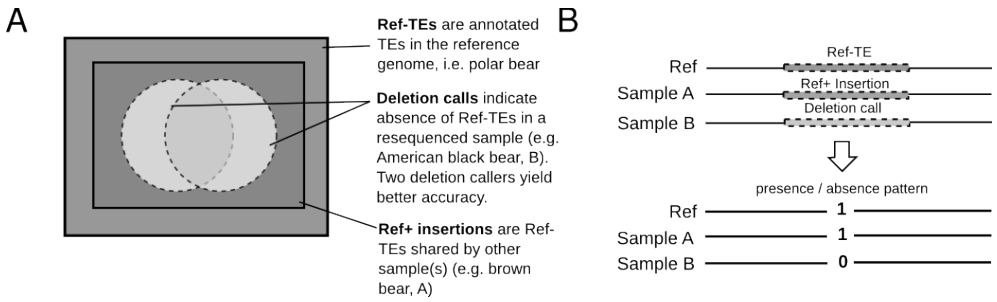


Supplementary Figure 1. Flowchart of the TeddyPi pipeline. Input data: TeddyPi requires a reference genome (Ref. Genome, FASTA format) and annotation of repetitive regions and assembly gaps (as BED files). TE and SV calls from resequenced samples, that were mapped against the reference genome are processed and data from multiple TE/SV callers (denoted as A and B) can be utilized (VCF files). TeddyPi is configured with a main configuration file that stores parameters and information on samples. Additional configuration files for each utilized TE/SV caller are needed and define the filtering steps applied to the data. The configuration files are read by all modules in the program routine of TeddyPi.

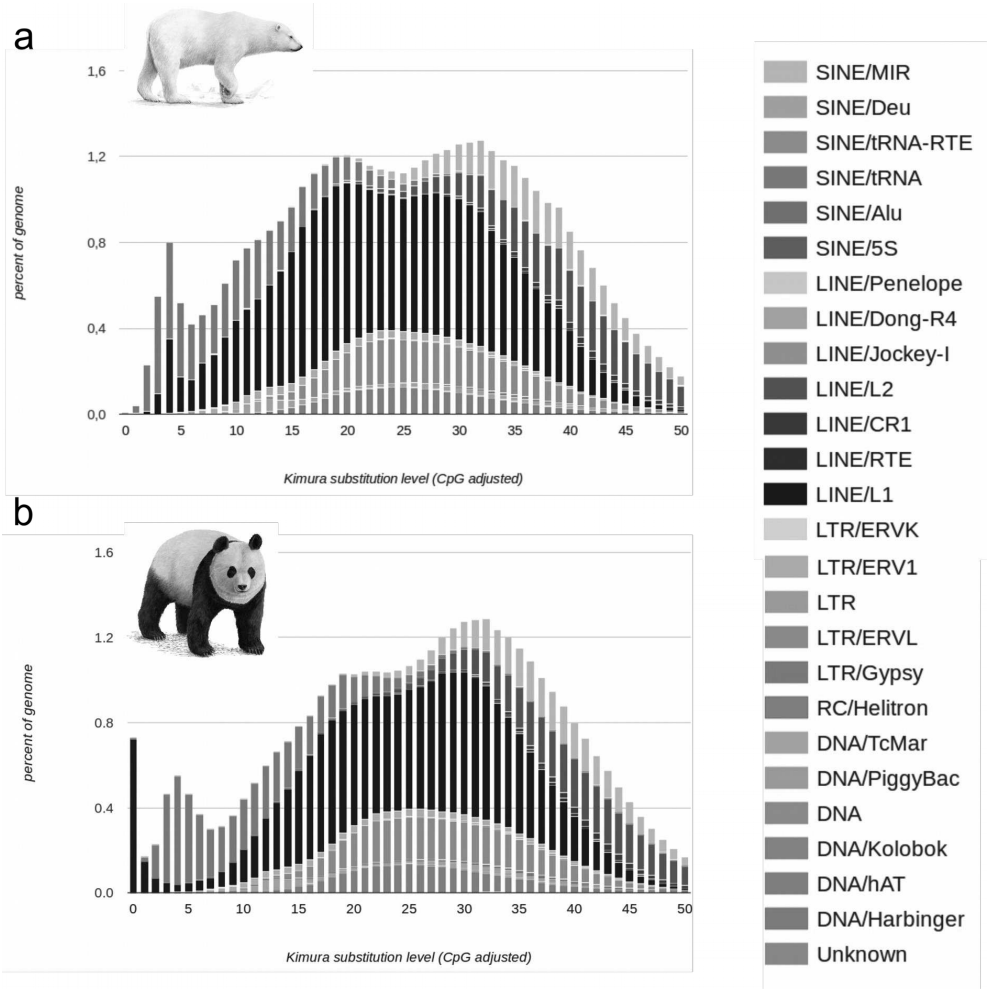
In the routine, first TE/SV callsets are filtered and if more than one are supplied all TE and all deletion (DEL) calls are integrated to a non-redundant set or by intersection (teddypi.py). The processing and filter methods are defined in tpi_filter.py (not shown). For all samples TE and DEL datasets are given as intermediate output. In tpi_ortho.py, data from all samples are integrated for TE and DEL calls respectively to create the Ref+ and Ref- datasets. Finally, both datasets are unified in tpi_unite.py and a presence/absence is stored in NEXUS format.



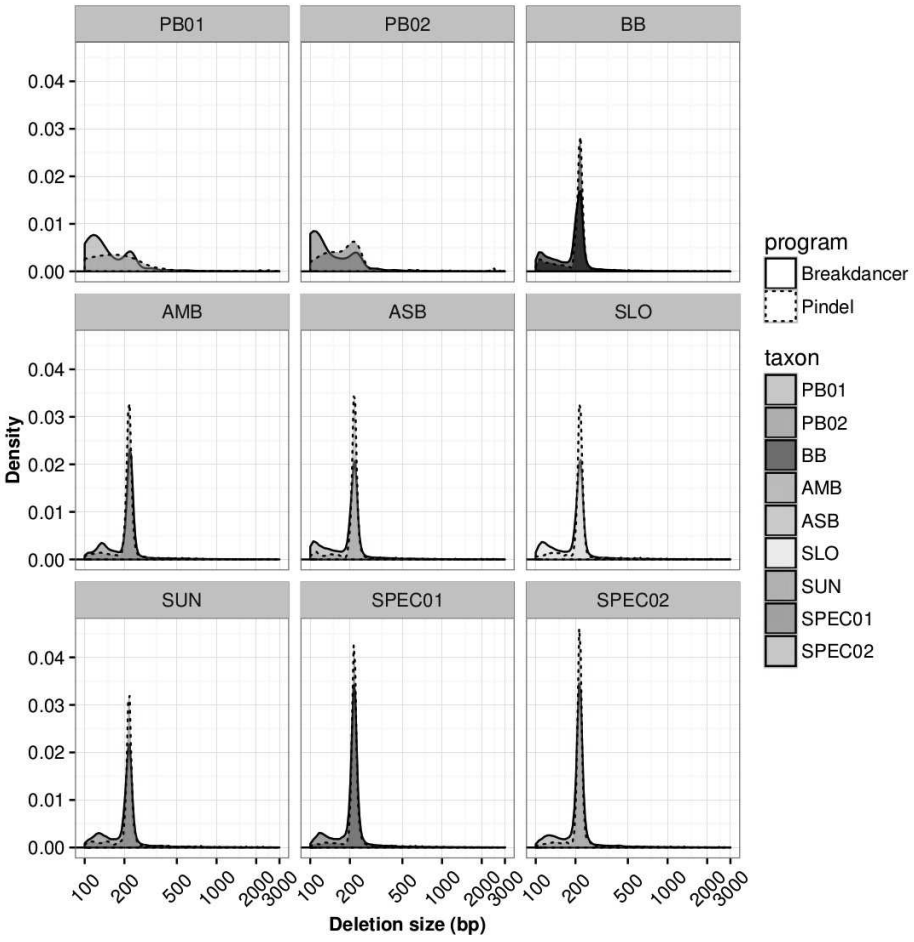
Supplementary Figure 2. Considerations for the choice of TE callers. Depending on the position of the reference genome (REF) within the expected species tree, insertions on different branches can be detected by non-reference (Ref-) insertion callers and/or by reference insertion-callers. In the different topologies A), B), C) the reference genome is nested inside the tree and not an outgroup. Ref- insertions (depicted as '-' at the branches) can therefore be detected on branches, that do not lead to the reference genome. The reference (Ref+) insertions (depicted as '+' along the branch) shared with additional taxa require Ref+-insertion calling to support the internode branches. D) In this example the reference genome is placed as the outgroup. Only in this case the sole use of a non-reference caller can find insertions supporting terminal and internode branches to taxon 1, 2 and 3. Depending on the initial phylogenetic hypothesis or knowledge about the taxa under study, a selection for Ref-, Ref+ or a combined detection approach needs to made.

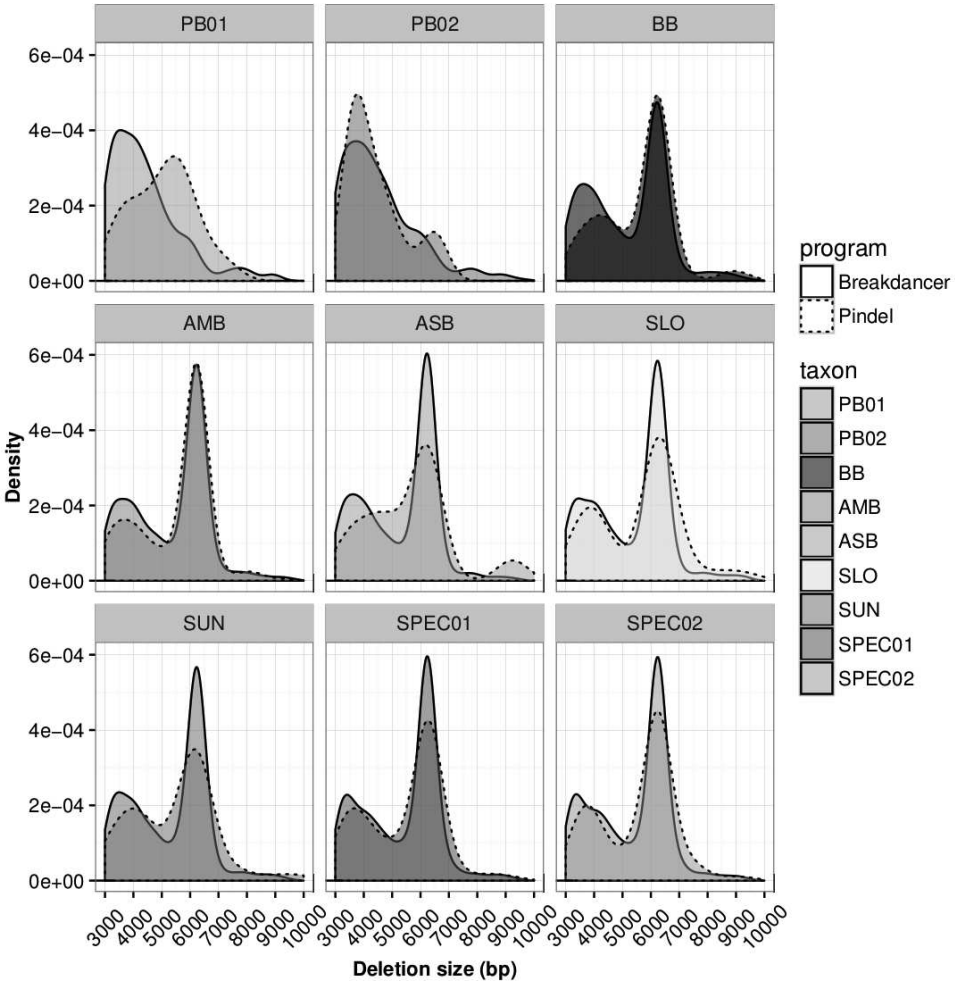


Supplementary Figure 3. The principle of deletion calling. A) TEs annotated in the reference genome (Ref-TEs, blue area) can either be private to the reference genome or shared with other species, if the insertion occurred in a common ancestor of both. TEs shared by the reference and other species are thus a subset of all Ref-TEs (grey area). To infer presence/absence patterns for such TEs, sample genomes are screened for deletion signatures with two programs (green circles). The presence of a deletion signature thereby indicates the absence of the TE insertion. Combining two sets of deletions called by two programs (here Pindel and Breakdancer) minimizes recognizing false positive Ref+ calls. B) A schematic alignment shows how the information of annotated reference TEs and the presence of a deletion call in sample B indicates a Ref+ insertion in sample A. If sequencing reads map normally and no deletion or any other structural variant is discovered by a TE or SV caller, the presence of the same TE in sample A and the reference assembly is inferred. For sample B at least one deletion call was made. Subsequently, the locus is recorded as 1-1-0 for the presence of TE insertions.

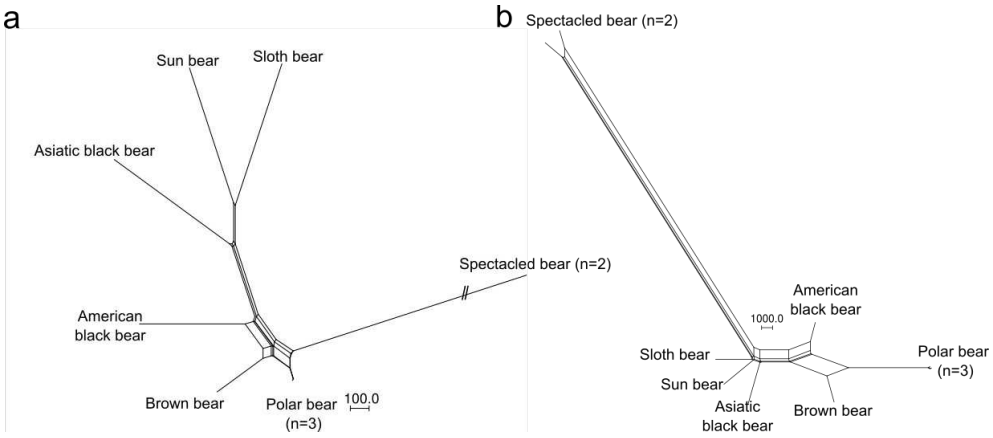


Supplementary Figure 4. Repeat landscapes from genome assemblies of polar bear (A) and giant panda (B). The graphs show the relative amount of each transposable element group in the genome in bins of 1% divergence to their consensus sequences. The divergence is shown on the x-axis and calculated as CpG adjusted Kimura-2-parameter substitutions to the consensus sequences. The y-axis shows the percentage of genome coverage for each TE. The repeat landscape for polar bear was generated with RepeatMasker. The repeat landscape for giant panda was copied from http://repeatmasker.org/species/ailMel.html [last accessed 2016/05/03].

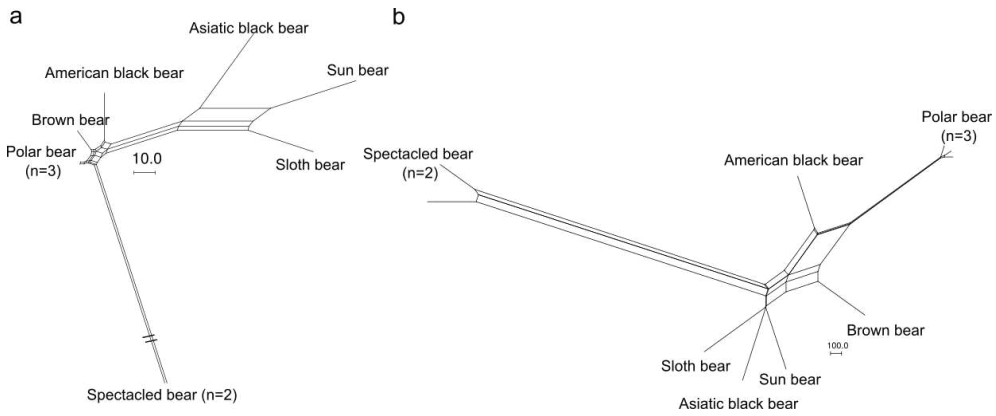

PB01
PB02
BB
AMB
ASB
SLO
SUN
SPEC01
SPEC02
Density
0.00
0.01
0.02
0.03
0.04
100
200
500
1000
2000
3000
Deletion size (bp)
program
Breakdancer
Pindel
taxon
PB01
PB02
BB
AMB
ASB
SLO
SUN
SPEC01
SPEC02



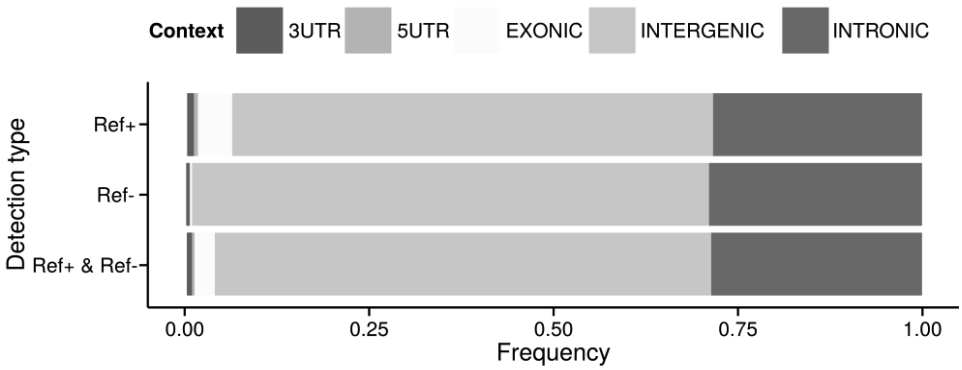
Supplementary Figure 6. Length distribution for deletions called by Pindel (dotted lines) and Breakdancer (solid lines) for each analyzed genome. The plots show the density for deletion of lengths between 3,000 bp and 10,000 kb. A peak around 6,000 bp indicates deletions originated by full length LINE-1 insertions. Shorter peaks likely originate from 5'truncated LINEs. The relative abundance of the 6,000 bp peak is similar in all samples, except polar bear, which exhibits a greater extent of deletion <6,000 bp. Sample names are polar bear (PB), brown bear (BB), American black bear (AMB), Asian black bear (ASB), sloth bear (SLO), sun bear (SUN), and spectacled bear (SPEC).



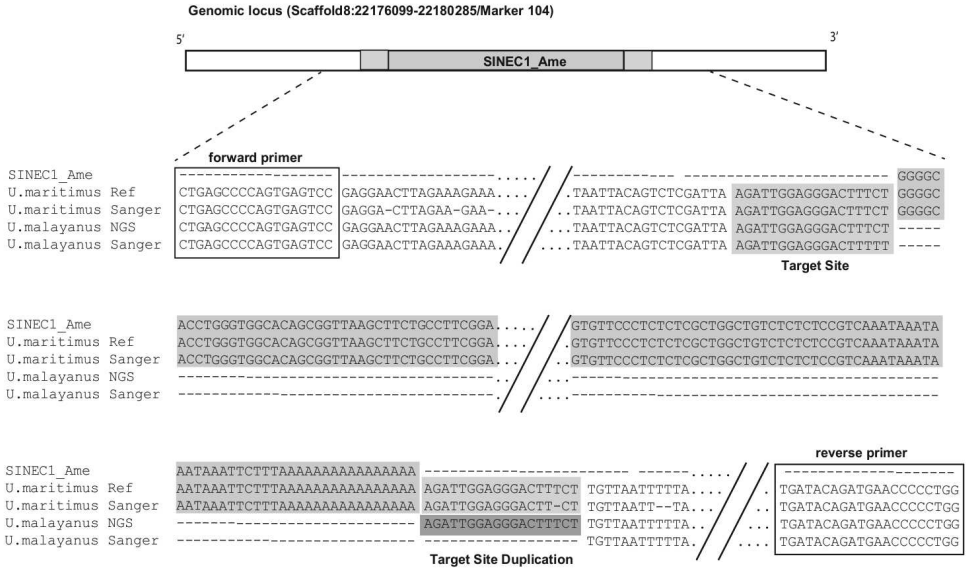
Supplementary Figure 7. Phylogenetic networks reconstructed from SINE insertions shown separately for Ref- insertions (A) and Ref+ insertions (B). A) Parsimony splits network from 61,026 Ref- SINE insertions have better resolution for the relationship between Asiatic black, sun and sloth bear than among polar bear, American black and brown bear. B) Parsimony splits network from 71,067 Ref+ SINE insertions resolve the relationship among polar bear, brown bear and American black bear. The edges between the three Asiatic bears are short and allow only limited resolution, but are consistent with the species tree.



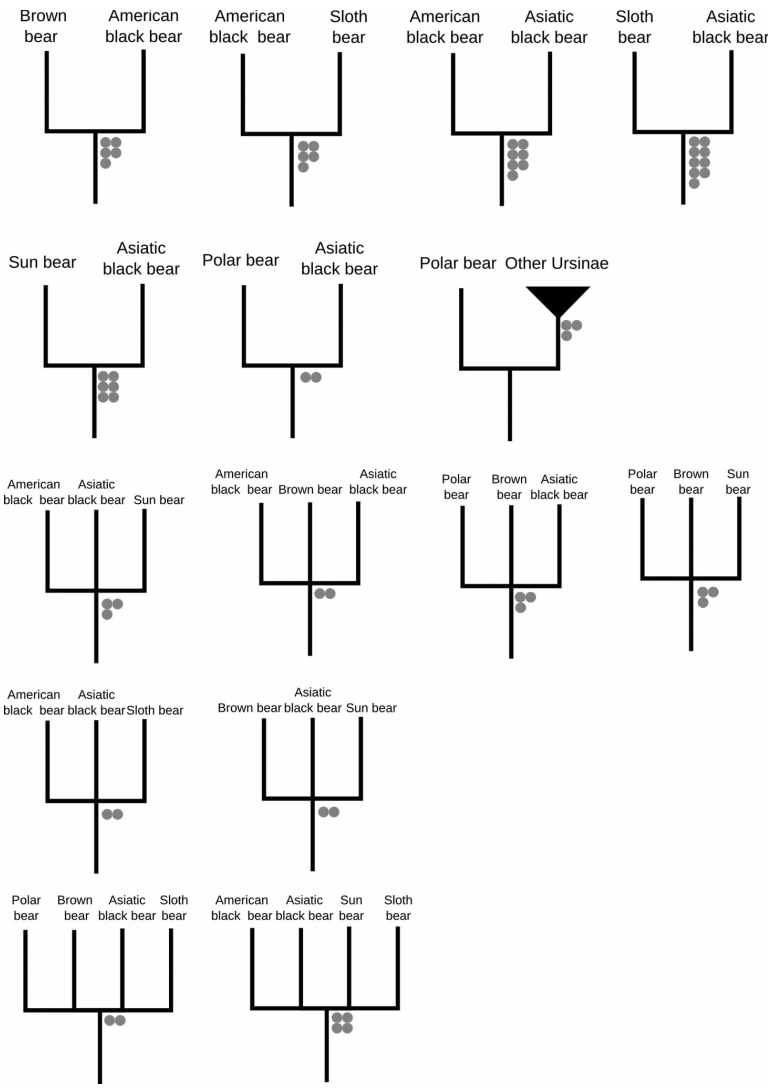
Supplementary Figure 8. Phylogenetic networks reconstructed from LINE1 insertions shown separately for Ref- insertions (A) and Ref+ insertions (B). A) A parsimony splits network from 6,455 LINE1 Ref- insertions with a minimum threshold of one character per branch. As for SINEs (Fig S5), Ref- insertions have better resolution for the relationship between Asiatic black, sun and sloth bear than for polar bear, brown bear and American black bear, respectively. The nested position of the reference genome used for TE calling causes the polar bear to appear in the center of the network. B) Parsimony splits network calculated from 11,965 LINE1 Ref+ insertions (threshold 1 character per edge) produce a long edge to the polar bear, brown bear plus American black bear, supporting a sister group relationship between them. The edges between the Asiatic bears are short and show only limited resolution from this type of marker.



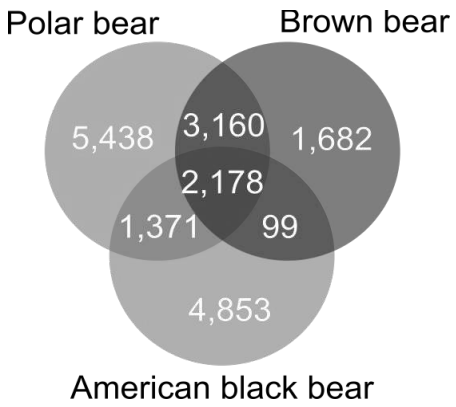
Supplementary Figure 9. Insertion frequency of TEs (SINEs and LINEs combined) into different genomic contexts in the polar bear genome. Color coding for genomic contexts is explained in the legend; the frequency describe the relative amount of TE insertions found in the respective genomic background. The insertion frequency is given separately for the different detection types: reference insertions (Ref+), non-reference insertion (Ref-) and the combined dataset. As expected, most insertions occurred in non-coding regions, i.e. intergenic regions and introns.



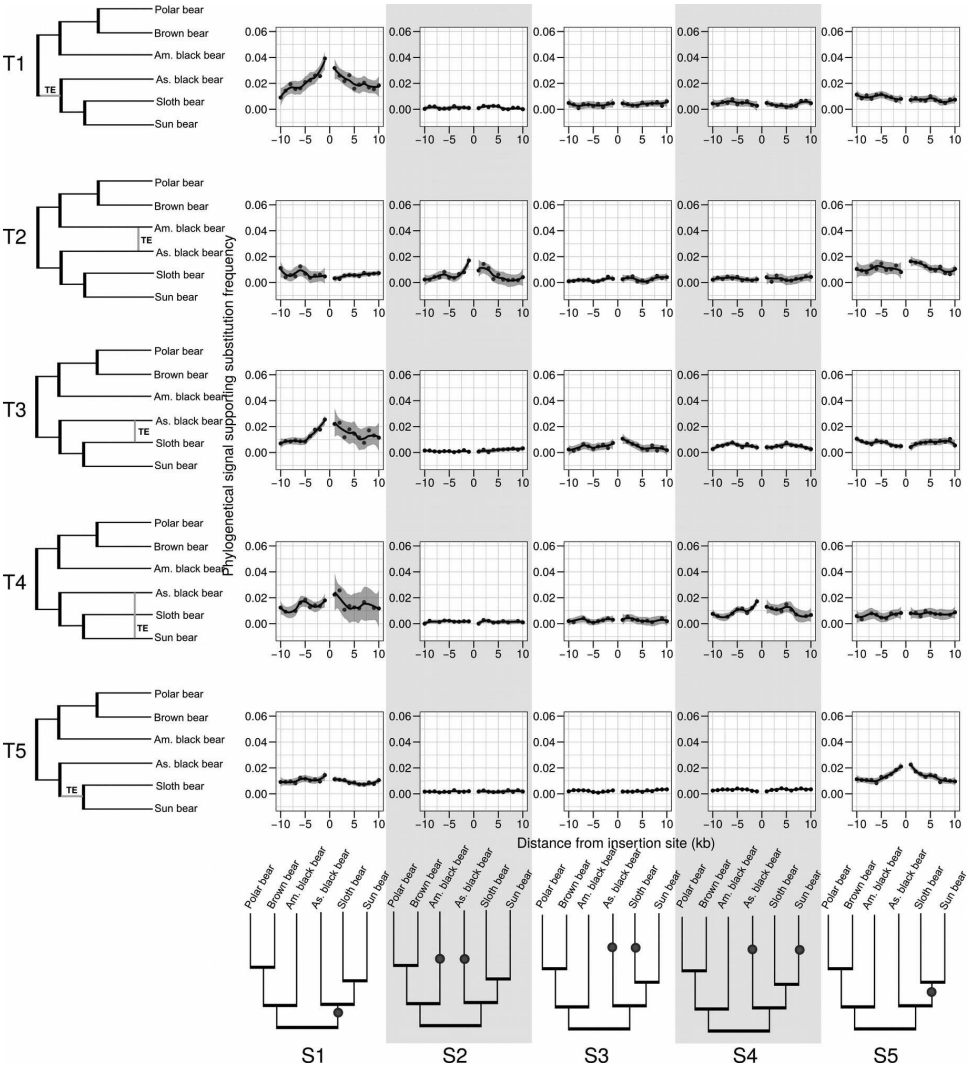
Supplementary Figure 10. Alignment of marker 104. Alignment of a genomic locus that harbors a reference SINEC1_Ame insertion in the polar bear (*U. maritimus*) while the insertion is absent in the sun bear (*U. malayanus*). The target site duplication (TSD) are highlighted and positions of conserved primers are boxed. For *U. maritimus* the genome assembly sequence (Ref) and the Sanger validated sequence are shown (Sanger), for *U. malayanus* the Illumina-based consensus (NGS) and Sanger sequences are shown. Note that, the consensus sequence in the Illumina sequence have a false target site duplication (TSD, highlighted in red). Sanger sequencing of the same sample used for Illumina sequencing revealed the absence of the SINE insertion and the TSD in *U. malayanus.*



Supplementary Figure 11. Phylogenetic signal from TE markers that are species-tree incongruent based on validation experiments. The topologies include phylogenetic signals that match phylogenetic hypotheses selected from the *in silico* predictions (e.g. shared TE insertions for Asiatic black and sloth bear, as shown in Table S12) and signals that contradict the *in silico* predictions, i.e. markers with partially erroneous predictions. Signals with only one supporting TE insertions are not shown graphically and listed in Data S1. TE insertions found in more than two species, are drawn on polytomic trees. Each red dot represents one TE insertion.



Supplementary Figure 12. Venn Diagram showing conflict among Polar bear, brown bear and American black bear on basis of inferred SINE insertions. The insertions numbers were extracted from the presence/absence table.



Supplementary Figure 14. Phylogenetic signals in the genomic sequences flanking the TEs. The panels show the frequencies of substitutions that support specific phylogenetic signals (specified in columns S1-S5) in windows of 1 kb surrounding the TE (from -10 to 10 kb) insertion site among loci carrying TE insertions for different phylogenetic signals (rows T1-T5). Each row (T1-T5) represents TE loci as indicated by the phylogenetic tree on the left-hand side (green branches). Vertical bars in trees T2-T4 connect branches in which the TE is present; they indicate TE insertions representing a phylogenetic signal incongruent to the species tree. The columns S1-S5 indicate substitutions with different phylogenetic signals. The signal is shown as blue dot in the trees in the bottom. For example, the first panel (T1-S1), show the frequency of substitutions that are shared by Asiatic black, sun and sloth bear (S1) in TE loci that show the same phylogenetic signal (T1). The second panel (T1-S2), shows the frequency of substitutions shared by American and Asiatic black bear in the same set of TE loci, as before (both are in row T1). For a detailed discussion refer to Supplementary Note 2.

Supplementary Table 1. List of genomes analyzed in this study.

Binomial name	Common name	ID	Accession number	Coverage	Insert size
Ursus maritimus	Polar bear[1]	PBREF	http://gigadb.org/dataset/100008	100X	
Ursus maritimus	Polar bear	PB01	SRR518686, SRR518687	11.8X	241 bp, SD= 19.8
Ursus maritimus	Polar bear	PB02	SRR518661, SRR518662	12.15X	267 bp, SD=31.3
Ursus arctos	Brown bear	BB	SRR935592, SRR935595, SRR935624, SRR935628	18.97X1	479 bp, SD= 22.6
Ursus americanus	American black bear	AMB	SRR518723	19.31X	300 bp, SD= 46.2
Ursus malayanus	Malayan Sun Bear	SUN	PRJEB9724	9.05	471 bp, SD= 32.4
Ursus ursinus	Sloth bear	SLO	PRJEB9724	9.09X	482 bp, SD= 23.5
Ursus thibetanus	Asiatic black bear	ASB	PRJEB9724	9.92X	482 bp, SD= 27.8
Tremarctos ornatus	Spectacled bear	SPE01	PRJEB9724	9.62X	476 bp, SD= 32.1
Tremarctos ornatus	Spectacled bear	SPE02	PRJEB9724	9.37X	474 bp, SD= 40.3

Note - [1]denotes the reference sequence, named polar bear genome; [2]SD = standard deviation

Supplementary Table 2. Selected phylogenetic hypotheses subject to validation experiments. The table shows the different taxon sets, for which synapomorphic TE loci were selected from the *in silico* data set. For each set, the range of primer IDs is given as well as a brief description for each hypothesis' origin. Hypotheses were based on the species tree or alternative phylogenetic trees proposed for Ursidae. For primers 120-179, loci were selected randomly for phylogenetically informative TE insertion from the specified datasets. Primer 1-29, were used to preliminary experiments and therefore not listed.

Primer ID	TE presence / Selection criterion	Description
30-39	Asiatic black bear, sun bear, and sloth bear	"Species tree"
40-49	Sloth bear and sun bear	"Species tree"
50-59	Sloth bear and American black bear	Autosomal / Y genes (Kutschera et al. 2014)
60-69	American black bear, Asiatic black bear, and sun bear	mtDNA tree
70-79	American black bear and Asiatic black bear	mtDNA tree
80-89	Sun bear and Asiatic black bear	Alternative topology from autosomal genes
90-99	Brown bear, American black bear, and polar bear	"Species tree"
100-119	Polar bear, brown bear, Asiatic black bear, and sun bear	
120-139	random Ref+	
140-149	random RetroSeq	
150-159	random Pindel	
160-169	random Breakdancer	
170-179	random RetroSeq + Mobster	

Supplementary Table 3. Repetitive elements in the polar bear genome sequence. The genome has been screened with RepeatMasker for Carnivore specific repeats in strict mode.

Element	Number of elements	Length (bp)	% of genome
SINEs	1,223,168	194,257,084	8.42
Alu/B1	0	0	0
MIRs	499,702	74,868,620	3.24
LINEs	978,888	492,525,513	21.34
LINE1	561,205	380,530,435	16.48
LINE2	350,263	96,987,569	4.20
L3/CR1	48,769	10,829,082	0.47
RTE	16,999	3,949,655	0.17
LTR	320,346	122,027,132	5.29
ERVL	92,885	40,217,869	1.74
ERVL-MaLRs	150,576	52,141,355	2.26
ERV_classI	50,572	22,967,428	0.99
ERV_classII	1,105	510,735	0.02
DNA	340,447	70,582,462	3.06
hAT-Charlie	196,435	37,647,590	1.63
TcMar-Tigger	50,119	14,904,611	0.65
Unclassified	6,539	1,107,006	0.05
Total interspersed repeats		880,499,197	38.14
Small RNA	751,454	121,127,388	5.25
Satellites	145	20,875	0
Simple repeats	632,864	28,020,132	1.21
Low complexity	103,265	5,288,810	0.23
Unmasked sequence			60.4

Supplementary Table 4. Prediction counts from RetroSeq SINE calls for raw calls and each filtering step. First, SINE insertion were selected from the dataset (SINEs). The call sets were filtered step-wise for homozygousity and quality (Quality), vicinity to assembly gaps (Filtered Gaps), vicinity to annotated TEs of the same type in the reference genome (Adjacent TEs) and per-sample defined coverage threshold per site (Coverage).

Sample	Raw	SINEs	Quality	Filtered Gaps	Adjacent TEs	Coverage
Polar bear 01	10,007	4,487	208	174	92	91
Polar bear 02	11,929	4,831	301	233	150	149
Brown bear	10,587	20,655	4,634	4,135	3,059	3,056
American black bear	59,728	38,193	11,417	10,750	7,374	7,372
Asiatic black bear	44,597	28,520	9,129	8,540	6,731	6,725
Sloth bear	45,223	28,886	13,871	13,196	10,705	10,697
Sun bear	48,792	32,458	12,875	12,285	10,015	10,005
Spectacled bear 01	95,630	72,520	36,914	36,205	29,654	29,638
Spectacled bear 02	95,878	73,103	36,659	35,950	29,433	29,422
Total	696,041	303,653	126,008	121,468	97,213	97,155

Supplementary Table 5. Predictions counts from Mobster for raw calls and each filtering step for SINEs. First, SINE insertions with at least 4 supporting reads on 5' and 3' end of the insertion were selected from the dataset (column SINEs). The call sets were filtered step-wise for quality (Quality), vicinity to assembly gaps (Filtered Gaps), vicinity to annotated TEs of the same type in the reference genome (Adjacent TEs) and per-sample defined coverage threshold per site (Coverage). Note that Mobster does not give zygosity information.

Sample	Raw	SINEs	Filtered Gaps	Adjacent TEs	Coverage
Polar bear 01	14,589	6,887	5,351	1,215	1,002
Polar bear 02	15,752	6,832	5,276	1,326	1,087
Brown bear	35,471	20,063	18,566	9,360	9,186
American black bear	73,865	47,542	44,857	14,701	14,439
Asiatic black bear	40,425	26,170	24,820	14,106	13,887
Sloth bear	41,772	28,074	26,672	15,530	15,362
Sun bear	47,494	32,905	31,432	17,117	16,833
Spectacled bear 01	84,380	66,838	65,297	40,466	40,221
Spectacled bear 02	88,578	70,278	68,652	40,725	40,509
Total	491,193	305,589	290,923	154,546	152,526

Supplementary Table 6. Prediction counts from RetroSeq LINE1 calls for raw calls and each filtering step. LINE1 insertions were selected from the dataset (L1s). The call sets were filtered step-wise for homozygosity and quality (Quality), vicinity to assembly gaps (Filtered Gaps), vicinity to annotated TEs of the same type in the reference genome (Adjacent TEs) and per-sample defined coverage threshold per site (Coverage).

Sample	LINE1	Quality	Filtered Gaps	Adjacent TEs	Coverage
Polar bear 01	5,271	197	114	37	36
Polar bear 02	6,814	336	170	58	56
Brown bear	18,446	3,859	2,088	1,004	1,000
American black bear	20,004	5,168	4,058	2,330	2,327
Asiatic black bear	15,117	5,069	3,318	1,919	1,912
Sloth bear	15,251	5,854	4,253	2,669	2,663
Sun bear	15,310	5,580	4,082	2,634	2,630
Spectacled bear 01	21,288	10,013	9,145	6,186	6,174
Spectacled bear 02	20,965	9,750	8,866	5,972	5,964
Total	138,466	45,826	36,094	22,809	22,762

Supplementary Table 7. Predictions counts from Mobster for LINE1 calls and each filtering step. First, SINE insertions with at least 4 supporting reads on 5' and 3' end of the insertion were selected from the dataset (column SINEs). The call sets were filtered step-wise for quality (Quality), vicinity to assembly gaps (Filtered Gaps), vicinity to annotated TEs of the same type in the reference genome (Adjacent TEs) and per-sample defined coverage threshold per site (Coverage). Note that Mobster does not differentiate between homo- and heterozygosity.

Sample	LINE1	Filtered Gaps	Adjacent TEs	Coverage
Polar bear 01	7,609	3,905	211	170
Polar bear 02	8,818	4,427	210	169
Brown bear	15,096	9,790	1,406	1,319
American black bear	25,394	20,003	1,751	1,691
Asiatic black bear	14,007	10,032	1,961	1,865
Sloth bear	13,504	9,903	2,053	1,989
Sun bear	14,338	10,628	2,303	2,213
Spectacled bear 01	17,184	15,029	4,382	4,256
Spectacled bear 02	17,892	15,683	4,415	4,297
Total	133,842	99,400	18,692	17,969

Supplementary Table 8. Summary of non-reference TE insertion counts in Ursinae for SINEs and LINEs with values from RetroSeq and Mobster and their overlap.

	SINEs			LINE1s		
Sample	RetroSeq	Mobster	Overlap	RetroSeq	Mobster	Overlap
Polar bear 01	91	1,002	65	36	170	8
Polar bear 02	149	1,087	120	56	169	14
Brown bear	3,056	9,186	2,518	1,000	1,319	221
American black bear	7,372	14,439	6,711	2,327	1,691	556
Asiatic black bear	6,725	13,887	5,727	1,912	1,865	729
Sloth bear	10,697	15,362	9,434	2,663	1,989	1,104
Sun bear	10,005	16,833	8,594	2,630	2,213	1,080
Spectacled bear 01	29,638	40,221	25,960	6,174	4,256	1,993
Spectacled bear 02	29,422	40,509	25,330	5,964	4,297	2,029
Total	97,155	152,526	84,462	22,762	17,969	7,734

Supplementary Table 9. Filtering results for the Breakdancer dataset. From the raw dataset, deletions were selected and filtered for a length >100 bp and < 10 kb (Size). Then, the call sets were filtered step-wise for vicinity to assembly gaps (Filtered Gaps) and for vicinity or overlaps with satellite DNA, and other repetitive sequences in polar bear reference genome that were not an interspersed repeat (Repeat-filtered). Finally, call sets were filtered for regions of extraordinary high coverage (Coverage).

				Filtered		Coverag
Sample	**Raw**	**Deletion**	**Size**	**Gaps**	**Repeat-filtered**	**e**
Polar bear 01	5,079	3,963	2,383	1,702	1,403	1,337
Polar bear 02	4,675	3,791	3,348	2,520	2,133	2,033
Brown bear	57,097	35,210	27,088	24,576	23,082	22,986
Am black bear	33,191	29,303	29,036	26,756	23,893	23,607
As black bear	69,487	43,393	38,800	35,939	33,620	33,465
Sloth bear	72,885	44,724	40,811	37,829	35,212	35,040
Sun bear	70,014	43,155	39,109	36,284	33,871	33,638
Spectacled bear 01	146,473	87,980	80,994	77,021	72,405	72,127
Spectacled bear 02	143,694	92,061	83,477	79,347	72,689	71,780
total	602,595	383,580	345,046	321,974	298,308	296,013

Supplementary Table 10. Filtering results for the Pindel dataset. From the raw dataset, deletions, homozygous deletions were selected and filtered for a length >100 bp and < 10 kb (Size). Then, the call sets were filtered step-wise for vicinity to assembly gaps (Filtered Gaps) and for vicinity or overlaps with satellite DNA, and other repetitive sequences in polar bear reference genome that were not an interspersed repeat (Repeat-filtered). Finally, call sets were filtered for regions of extraordinary high coverage (Coverage).

Sample	Raw	Deletion	Size	Filtered Gaps	Repeats-filter	Coverage
Polar bear 01	155,489	8,842	111	75	54	43
Polar bear 02	157,895	8,932	124	80	64	54
Brown bear	737,502	84,769	2,561	2,311	1,910	1,863
Am black bear	1,031,231	204,739	5,503	5,145	4,312	4,092
As black bear	758,266	74,660	1,213	1,103	935	883
Sloth bear	829,500	96,421	1,159	1,054	877	805
Sun bear	785,786	87,122	1,060	970	805	733
Spectacled bear 01	1,357,194	236,097	3,207	3,027	2,603	2,344
Spectacled bear 02	1,320,685	213,305	2,776	2,613	2,273	2,048
Total	11,746,167	1,014,887	17,714	16,378	13,833	12,865

Supplementary Table 11. Results of Ref+ insertion processing. Deletion calls from Breakdancer and Pindel were combined to a non-redundant set (DEL_nr) and screened for intersection with TEs in the polar bear reference genome (TEs), from which calls corresponding to SINE and LINE1 insertions were extracted (SINEs, LINE1s). Other deletions corresponding to TE insertions are counted as 'Other'.

Sample	DEL_nr	TEs	SINEs	LINE1s	LINE1_frag	LINE1 >5kb	Other
Polar bear 01	1,324	911	535	251	246	5	125
Polar bear 02	2,005	1,350	723	422	414	7	205
Brown bear	22,976	19,862	14,157	3,004	2,945	59	2,701
Am black bear	23,481	20,877	16,247	2,900	2,829	71	1,730
As black bear	33,458	30,195	22,335	3,954	3,851	103	3,906
Sloth bear	35,029	31,645	23,726	3,998	3,897	101	3,921
Sun bear	33,625	30,563	23,026	3,666	5,568	98	3,871
Spectacled bear 01	72,207	68,181	54,917	6,154	6,000	154	7,110
Spectacled bear 02	71,329	67,105	55,333	6,260	6,105	155	5,512
Total	295,434	270,689	210,999	30,609	29,855	754	29,081

Supplementary Table 12. Heterozygous loci identified by PCR. For each species with >=1 heterozygous PCR amplicons, the number of heterozygous (Het) and homozygous (Hom) loci are indicated. Heterozygosity (% Het) is estimated by dividing the number of heterozygous amplicons by total amplicon count.

Species	Het	Hom	Total	% Het
Polar bear	2	49	51	3.92
Brown bear	8	39	47	17.02
American black bear	3	48	51	5.88
Asiatic black bear	4	67	71	5.63
Sun bear	3	54	57	5.26
Sloth bear	1	55	56	1.79
Total	21	312	333	6.31

Supplementary Table 13. KKSC test results for SINE insertion counts. The KKSC test was performed on two clades, consisting of species triplets. The first clade (PB-BB-AMB) is polar bear, brown bear and American black bear. The most likely tree according to the test is ((Polar bear, Brown bear), American black bear) at p=1.6106E-207. Gene flow is inferred from polar bear and American black bear to brown bear at p=3.2169E-193. The second clade (ASB-SUN-SLO) is Asiatic black, sun and sloth bear and the most likely topology is ((Sun bear, Sloth bear), Asiatic black bear) and hybridization is rejected at p=0.9686.

Clade	Test type	Significance level	Test values	Critical border at p<=0.05	Critical border at p<=0.01
PB-BB-AMB	hybridization test	8.8623E-287	1371 vs 1371 + 99	76	100
	tree test	1.0404E-159	3160 vs 1371	133	174
ASB-SUN-SLO	hybridization test	0.6066	278 vs 278 + 265	47	61
	tree test	6.8845E-47	3993 vs 2809	163	213

Supplementary Data

Supplementary Data 1. Spreadsheet with loci and primer sequences

Supplementary Data 2. Tab-separated file of final TE dataset

Supplementary Data 3. FASTA alignments of selected validated loci (ZIP archive)

Supplementary Notes

Supplementary Note 1- Discrepancies between NGS-generated and Sanger sequences.

The paired-end Illumina sequences of each genome were mapped against the polar bear reference genome, then SNVs were called from the short-read alignments to generate consensus sequences of the resequenced genomes (Kumar et al. 2016). In the consensus sequences, Ref- TE insertions are generally not assembled. The presence of a TSD, flanking the potential TE insertion, was frequently observed in the consensus sequence. Experimental validation using Sanger sequencing of the DNA from the same individual showed that the TSDs flanking the breakpoint were artificially generated during the mapping process and represent artifacts. The presence of TSDs in consensus sequences generated from short read alignments, is therefore not informative regarding the presence of a TE insertions in the resequenced genome. Accordingly, there was also no correlation between the TE prediction calls and the incorrectly generated TSDs in the consensus sequences (**Supplementary Fig. 10**).

Supplementary Note 2. Remarks on flanking sequence analysis.

In addition to test the association between TE insertion loci with flanking substitutions that exhibit the same phylogenetic signal, we screened the flanks for the presence of substitutions that support alternative phylogenetic hypotheses. We extracted flanking sequences shared by:

- S1: Asiatic black bear, Sun bear, and Sloth bear
- S2: Asiatic black bear and American black bear
- S3: Asiatic Black bear & Sloth bear S4: Asiatic Black bear & Sun bear
- S5: Sun bear & Sloth bear

The trees per row (T1-T5) show the selected topologies supported by TE insertions (**Supplementary Fig 13**). In the diagrams, the mean frequency of substitutions supporting phylogenetic signals are shown in 1 kb windows. The columns S1-S5 indicate for which taxa synapomorphic substitutions were selected for the frequency analysis. The blue dots in the trees S1-S5 in which taxa the substitutions are present. As described in the main text, substitutions supporting the phylogeny indicated by TEs accompany a TE insertion loci at different spatial scales (**Supplementary Fig 13 panels in the diagonal**). In general, we observe no substitutions supporting other topologies than indicated by the TE insertion. Interestingly, loci carrying TE insertions in Asiatic black bear and either sun or sloth bear show - in addition to phylogenetically congruent substitutions - elevated substitutions that group together all three Asian bear species (**Supplementary Figure 13 panel T3-S1, T4-S1**). A possible evolutionary model (model 1) to explain this pattern is that species-tree congruent TE loci reflect the speciation history of Asiatic black bear, sun and sloth bear. Introgressive hybridization post-speciation transferred the TE-containing loci between Asiatic black and one of its sister species. A second explanation (model 2) is that the TE insertion is an ancestral polymorphism cannot be ruled out for these cases. However, if the higher amount of substitutions supporting the species-tree phylogeny is considered a consequence of an older common ancestry (i.e. more time has past to accumulate substitutions) rather the first model is supported.

Other deviations between TE insertion and the surrounding substitutions can be observed in panels T1-S5 and T2-S5 (**Supplementary Fig. 13**). In panel T1-S5, in addition to the TE congruent substitutions that are shared by all three Asiatic bear species, an increased number of substitutions supporting the monophyly of sun and sloth bear are found. However, this pattern is expected as it reflects the speciation history of all three species.

In panel T2-S5, loci with TE insertions in American and Asiatic black bear show substitutions supporting also the monophyly of sun and sloth bear. This can be explained by the speciation history where the Asiatic black bear diverged from the ancestor of all three Asian bear species and subsequently exchanged alleles with a lineage related to the American black bear or maintained ancestral polymorphisms from the initial ursine radiation.

4.2 Whole-genome sequencing of the blue whale and other rorquals finds signatures for introgressive gene flow

Ulfur Arnason*, **Fritjof Lammers***, Vikas Kumar, Maria Nilsson & Axel Janke (*shared first authors)

Published in
Science Advances 4: eaap9873 (2018).
10.1126/sciadv.aap9873

Erklärung zu den Autorenanteilen an der Publikation
Was hat der Promovierende bzw. was haben die Koautoren beigetragen?

(1) zu Entwicklung und Planung
FL 50%; UA,AJ,MAN 50 %

(2) zur Durchführung der einzelnen Untersuchungen und Experimente
NGS-Datenverarbeitung: FL, VK (100%)

(3) zur Erstellung der Datensammlung und Abbildungen
Probenbeschaffung: UA (10 %); Probenorganisation, Zellkultur: UA, AJ, MAN (10%); DNA-Extraktion: MAN, AJ (10%); Erstellung der Abbildungen & Tabellen: FL (70%)

(4) zur Analyse und Interpretation der Daten
Genomeanalysen + Mapping: FL, VK (20 %); Erstellung Phylogenomischer Datensatz: FL (20%); Phylogenomische Analysen, de novo Assembly, PSMC Analysen, Divergenzzeitberechnung: FL 40 %; Interpretation der Ergebnisse: FL, MAN, AJ (20%)

(5) zum Verfassen des Manuskripts
FL 65 %, MAN: 10 %, AJ: 15%, UA: 10%

Zustimmende Bestätigungen der oben genannten Angaben:

_____ , _____	
Datum, Ort	Unterschrift Promovierender
_____ , _____	
Datum, Ort	Unterschrift Betreuer
_____ , _____	
Datum, Ort	ggfs. Unterschrift *corresponding author*

SCIENCE ADVANCES | RESEARCH ARTICLE

EVOLUTIONARY BIOLOGY

Whole-genome sequencing of the blue whale and other rorquals finds signatures for introgressive gene flow

Úlfur Árnason,[1]* Fritjof Lammers,[2,3,4]* Vikas Kumar,[2] Maria A. Nilsson,[2] Axel Janke[2,3,4]†

Copyright © 2018 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution NonCommercial License 4.0 (CC BY-NC).

Reconstructing the evolution of baleen whales (Mysticeti) has been problematic because morphological and genetic analyses have produced different scenarios. This might be caused by genomic admixture that may have taken place among some rorquals. We present the genomes of six whales, including the blue whale (*Balaenoptera musculus*), to reconstruct a species tree of baleen whales and to identify phylogenetic conflicts. Evolutionary multilocus analyses of 34,192 genome fragments reveal a fast radiation of rorquals at 10.5 to 7.5 million years ago coinciding with oceanic circulation shifts. The evolutionarily enigmatic gray whale (*Eschrichtius robustus*) is placed among rorquals, and the blue whale genome shows a high degree of heterozygosity. The nearly equal frequency of conflicting gene trees suggests that speciation of rorqual evolution occurred under gene flow, which is best depicted by evolutionary networks. Especially in marine environments, sympatric speciation might be common; our results raise questions about how genetic divergence can be established.

INTRODUCTION

Baleen whales (Mysticeti) are strikingly derived marine mammals that encompass the largest animals living on Earth (*1*); however, their evolution is only poorly understood. Today, 15 species of extant baleen whales are known, and the fossil record includes many additional extinct species (*2*). The gigantic blue whale (*Balaenoptera musculus*) with a length of 30 m and a weight exceeding 150 metric tons and the fin whale (*Balaenoptera physalus*) are the largest animals on Earth (*1*). Both belong to the rorqual family (Balaenopteridae). Baleen whales have undergone significant adaptations to marine life and are characterized by their lack of teeth, which have been replaced by keratin bristles, the baleen that is used for filter feeding (*3*). It has been estimated that the blue whale takes in up to 3.6 metric tons of krill every day to supply the energy demand of their huge body sizes (*3*). The large body size of whales allowed them to occupy novel ecological niches by enabling deep dives and to endure long periods of starvation to reach feeding grounds (*4*). The evolutionary history of baleen whales is debated, despite extensive analyses of molecular and morphological characteristics (*2*, *5*). Moreover, molecular analyses of baleen whale evolution disagree with each other depending on the applied marker and type of phylogenetic analysis (*5*–*8*). Of particular interest are the humpback whales (*Megaptera novaeangliae*) and gray whales (*Eschrichtius robustus*), which are each placed in a separate genus or even in its own family, mainly based on analyses of their derived anatomy (*1*). However, these classifications are not supported by recent molecular studies, which suggest that they evolved from within rorquals, making the latter paraphyletic. To reflect this discordance, we will use the family name Balaenopteridae sensu lato, that is, including Balaenopteridae and Eschrichtiidae.

It is difficult to envision that the baleen whales evolved by allopatric speciation under vicariance because the marine environment largely lacks physical barriers for mobile species like whales (*1*, *9*). The study of the evolution of whales is further complicated by the fact that whales can hybridize. In the case of the blue whale and the fin whale, genetic analyses have shown that the female hybrid carried a fetus and had mated with a blue whale (*10*). Thus, these two species, as well as other rorquals, may not be entirely reproductively isolated. In addition, rorquals have a conserved karyotype of $2n = 44$ chromosomes and an identical chromosomal C-banding pattern, which facilitate producing fertile offspring (*11*).

Genomic analyses allow detailed insight into evolutionary processes such as speciation or past hybridization events (*12*) and permit examination of long-standing evolutionary questions (*13*). Introgressive hybridization, speciation with gene flow, and incomplete lineage sorting (ILS) may cause different local genealogies across the genome that can be detected by analyzing whole-genome sequences (*14*). Compared to terrestrial species, genomic data are limited for marine mammals, and before this study, genomic data were only available for three baleen whales: the bowhead whale (*Balaena mysticetus*), the minke whale (*Balaenoptera acutorostrata*), and the fin whale (*15*, *16*).

Here, we present genomic data of six mysticete species including the humpback and gray whale and the largest extant animal ever lived, the blue whale. The data are analyzed under the multispecies coalescent (MSC) that incorporates the genome-wide heterogeneity of gene trees to accurately infer speciation history (*14*). In addition, the genomes allow us to study signals of recent and ancestral introgression, to place divergences into a solid temporal context, and to explore genetic diversity and past demographic history of baleen whales.

RESULTS

Genome sequencing and assembly

Genomic DNA from six baleen whales and a hippopotamus (*Hippopotamus amphibius*) were sequenced with Illumina technology. Reference genome mapping of the whale genome data against the bowhead whale genome (*16*) yielded genome coverages of 6.3 to 27.2× (table S1). RepeatMasker (*17*) identified 40.3% repetitive sequences in the bowhead whale genome assembly. Of these, 6 and 18% were short and long interspersed elements (SINE and LINEs), respectively (table S2). Except for the genomic fraction of SINEs, these results are consistent with the original analyses of Keane *et al.* (*16*). We identified, on average, 25 million fixed single-nucleotide differences relative to the bowhead whale genome (table S3). Consensus sequences of all baleen whale genomes were aligned per scaffold, and repetitive sequences, gaps, and ambiguous bases were

[1]Department of Brain Surgery, Faculty of Medicine, University of Lund, Lund, Sweden. [2]Senckenberg Biodiversity and Climate Research Centre, Senckenberg Gesellschaft für Naturforschung, Senckenberganlage 25, 60325 Frankfurt am Main, Germany. [3]Goethe University Frankfurt, Institute for Ecology, Evolution and Diversity, Biologicum, Max-von-Laue-Straße 13, 60439 Frankfurt am Main, Germany. [4]LOEWE Centre for Translational Biodiversity Genomics, Senckenberganlage 25, 60325 Frankfurt, Germany.
*These authors contributed equally to this work.
†Corresponding author. Email: axel.janke@senckenberg.de

Downloaded from http://advances.sciencemag.org/ on April 18, 2018

removed. Empirical analyses and simulations using the approximate unbiased (AU) test (*18*) showed that 20–kilo–base pair (kbp) genome sequence alignments contain sufficient information for statistically significant maximum likelihood (ML) gene tree inference (figs. S1 to S3). The aligned scaffolds yielded 34,192 genome fragments (GFs), each 20 kbp long, totaling 643,840 kbp for each whale. This represents 49% of the nonrepetitive genome sequence. Sequencing the hippopotamus genome yielded 1,684,446,285 filtered reads and a sequencing depth of 55× (table S4). The reads were assembled de novo with Minia (*19*) and scaffolded with SSPACE, resulting in a genome assembly of 2.43 Gbp with a scaffold N50 of 120 kbp. AUGUSTUS (*20*) identified 29,998 coding sequences (CDSs); 37.0% of the genome were masked as repetitive (table S5).

The evolution of whales

Model testing identified the generalized time-reversible model with gamma-distributed rate variation with invariable sites (GTR + 4G + I) as the best-fitting nucleotide substitution model for the ML analyses of GFs. An MSC species tree of baleen whales based on 34,192 GF trees was supported with posterior probabilities of 1.0 for all branches (Fig. 1A and fig. S4). The topology conforms to previous nuclear gene and mitochondrial DNA (mtDNA) analyses (*5*, *21*) and a Bayesian phylogeny of the mtDNA sequences reported herein (fig. S5). The primary characteristic of the tree is the clear distinction between the Balaenidae (right whales) and the branch harboring the five rorquals plus the gray whale (Balaenopteridae sensu lato). The humpback whale (genus *Megaptera*) groups within the rorquals, resulting in a paraphyly of the current genus *Balaenoptera*. The gray whale of the monotypic family Eschrichtiidae is placed inside rorquals as a sister lineage to fin and humpback whale. However, quartet scores, that is, the support for any of three possible phylogenetic arrangements around an internal branch, identified conflict in resolving the branch leading to the ancestor of the gray, fin, and humpback whale (Fig. 1A, branch no. 3). The three possible topologies for this branch receive similar quartet scores (Fig. 1B), contrasting to a posterior probability of 1.0. Thus, we find highly similar support for placing the gray whale as a sister group to blue whales and sei whales or as a distinct clade outside the blue/sei/fin/humpback whale cluster. Somewhat inconclusive support also marks the first branch inside rorquals (Fig. 1B, branch no. 2) that places the minke whale as a sister lineage to the remaining Balaenopteridae sensu lato with a quartet score of 0.7. Phylogenetic conflict is also present in a CONSENSE (*22*) analysis of the GF trees. Although a majority-rule consensus tree confirms the coalescent-based species tree (Fig. 1A and fig. S6), two alternative phylogenetic positions of the gray whale are equally strongly represented (table S6).

The position of the gray whale in the species tree is supported by 10,315 (30.2%) GF trees compared to 8918 (26.1%) and 8721 (25.6%) GF trees, which place the gray whale in different positions inside rorquals.

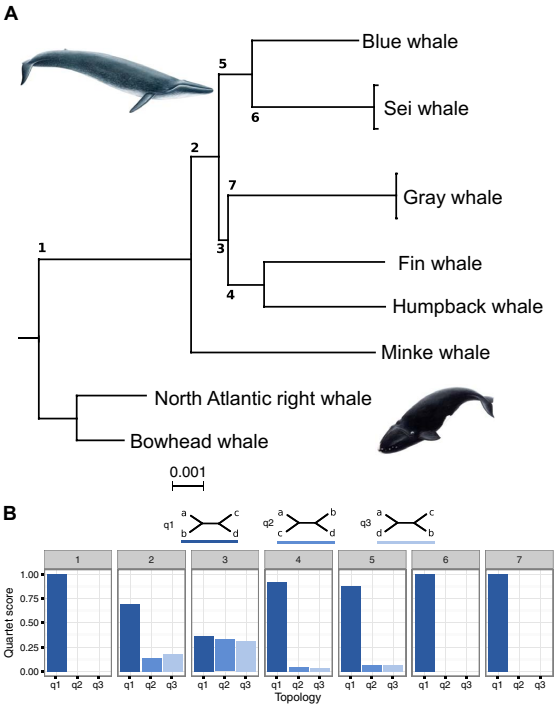


Fig. 1. MSC tree. (**A**) An MSC species tree was constructed from 34,192 individual GFs. Internal branches within Balaenopteridae are numbered 1 to 7. All branches receive maximal support ($P = 1.0$, ASTRAL analysis). Branch lengths were calculated from an ML analysis. Gray whales, family Eschrichtiidae, are placed inside Balaenopteridae as a sister group to fin and humpback whales. (**B**) ASTRAL quartet-score analyses for branches 1 to 7 (A). Quartet scores were calculated for the three possible arrangements (q1 to q3) for the respective branch. The principal quartet trees are depicted, with q1 representing the species tree. Branch nos. 2 and 3 receive only limited quartet scores, and no quartet can be significantly rejected.

Downloaded from http://advances.sciencemag.org/ on April 18, 2018

A placement of the gray whale outside rorquals is supported by 3507 GF trees (10.3%). A consensus network analysis (*23*) of the GF trees yields a large cuboid structure of connecting alternative branches in the center of the network that indicates conflicting signals for the position of the gray whale inside rorquals (Fig. 2). At a threshold for conflicting edges of 11%, the grouping of the humpback and fin whale, the sei and blue whale, and the bowhead and North Atlantic right whale is unambiguous. At lower thresholds, the phylogenetic signal becomes more complex, indicating additional phylogenetic conflict in the data (fig. S7).

Gene flow analyses

D statistic (*24*) and D_{FOIL} (*25*) analyses identified several gene flow signals among rorquals (Fig. 3A and data S1 and S2). We find significant gene flow signals between minke whale and the ancestors of the blue and sei whale and those of the fin and humpback whale, respectively. The D_{FOIL} analyses find a strong signal for gene flow between the ancestor of the blue and sei whale and the ancestor of the fin and humpback whale, which is likely a phylogenetic signal related to a placement of placing the gray whale into different positions (Fig. 3A and data S1 and S2). In addition, signal for recent gene flow was inferred reciprocally from the blue whale to the fin and humpback whale for about 1 to 1.5% of the genome. The *D* statistic analyses also identified numerous signals for gene flow between the ancestor of the blue/sei whale and gray whale and that of the humpback whale and gray whale. Note that the *D* statistic and D_{FOIL} analyses depend on the species tree as in Fig. 1A and the signal may vary for other constellations. Our interpretation, therefore, focuses on signals that are independent of the evolutionary placement of gray whales.

In addition to character-based parsimony analysis, gene flow may preferably be studied by topology-based ML analysis using PhyloNet (*26*). PhyloNet identifies a statistically significant signal for gene flow between the minke whale and the ancestor of the other rorquals (Fig. 3B). With equal likelihood probability, gene flow occurred from the ancestor of the humpback and fin whale to that of the minke whale (Fig. 3C). Furthermore, with a topology change of the gray whale as a sister group to blue and sei whale, gene flow occurs from the ancestor of the blue and sei whale to that of the minke whale (Fig. 3D). Each of the three reticulations shows inheritance probabilities of about 33%, resembling the quartet-score distribution of the coalescent tree analyses (Fig. 1B).

Genetic diversity and population size history

Genome-wide heterozygosity varies considerably among baleen whales (Fig. 4A and fig. S8). At approximately 5×10^{-4} heterozygous sites per nucleotide, estimates were lowest for the gray whale, the minke whale, and the two sei whales. The blue whale genome shows the highest

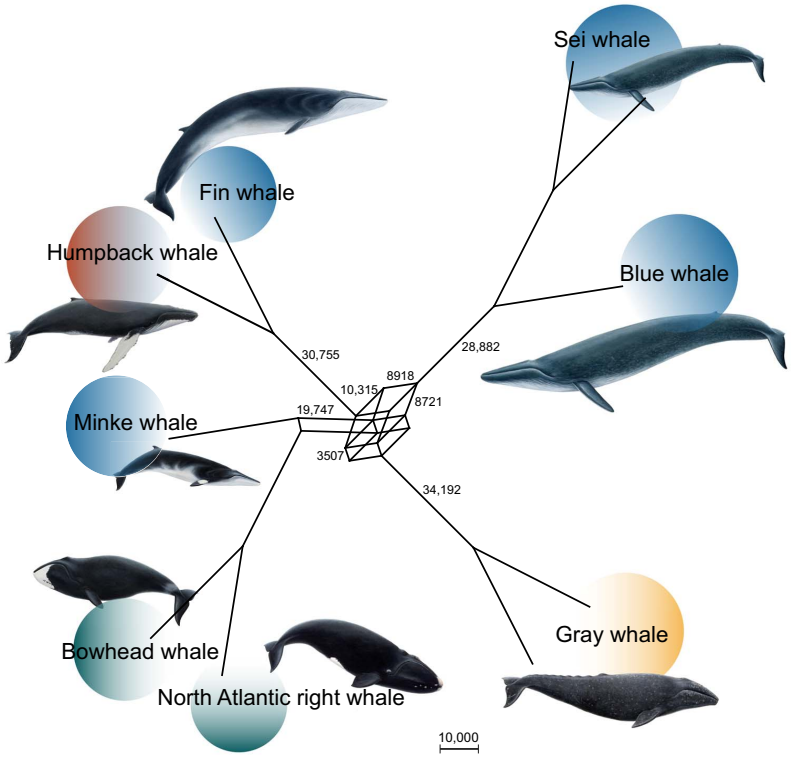


Fig. 2. Median network of 34,192 GF ML trees with 11% threshold. Conflicting evolutionary signals characterize the center of the network, which is equivalent to branch no. 3 in the species tree (Fig. 1). In addition, placing the minke whale has some conflicting signal, but the elongated rectangle indicates a higher degree of resolution. The number of supporting GFs is shown for selected splits. Colored circles indicate taxonomic classification. Blue, *Balaenoptera*; red, *Megaptera*; yellow, *Eschrichtius*; green, *Balaena* and *Eubalaena*.

Downloaded from http://advances.sciencemag.org/ on April 18, 2018

degree of heterozygosity, which is elevated even when compared to other mammals (*27*). Estimates for heterozygosity in downsampled genomic data of blue whale were similar, minimizing the effects of potential artifacts by higher sequence coverage (fig. S9). The history of the effective population size (N_e) over the last 5 million years (Ma) was modeled from the distribution of heterozygous sites across the genome using a pairwise sequentially Markovian coalescent (PSMC) (*28*) analysis (Fig. 4B and fig. S10). Ancestral effective population sizes for all baleen whales, particularly the large blue, fin, and humpback whales, were notably higher during the Plio-Pleistocene transition (PPT; 2.6 Ma ago) than recent estimates (Fig. 4B). After the mid-Pleistocene transition (MPT), N_e of most baleen whales was relatively stable, until approximately 100 thousand years (ka) ago, the time of the last interglacial. After this time, baleen whale populations decreased. In contrast, gray whale population size remained stable during the interglacial, and its population size even increased in more recent times. The blue whale maintained a larger population size than other whales, but their numbers decreased at 400 ka ago after the MPT. The minke and fin whale population increased somewhat at 200 to 300 ka ago, followed by a steady decline. The N_e of the humpback whale was rather constant since 1 Ma ago and then shows a decline by two-thirds of its population at some 30 ka ago. Our estimates of historical population sizes of the fin and minke whale are consistent with previous analyses (*15*).

Divergence time estimates

The phylogenomic reconstruction of a paraphyletic position of Cetacea among Artiodactyla and the placement of the Hippopotamidae are, for the first time, supported by genomic sequence data analyses (Fig. 5). The divergence times are based on five calibration points (table S8). Hippopotamidae diverged at 53.5 Ma ago, close to the appearance of archaeocetes in the fossil record at 50 Ma ago (*29*). Rorquals diverged in the late Miocene, between 10.48 and 4.98 Ma ago (table S9). The divergence time between baleen and toothed whales at 30.5 Ma ago coincides with the Eocene/Oligocene transition at 33 Ma ago (*30*), which probably triggered the radiation of modern whales.

DISCUSSION

Our genome analyses have shown that the evolution of Balaenopteridae sensu lato (hereafter referred to as rorquals) is not characterized by an ordered dichotomous divergence of lineages as would be expected with respect to speciation in most other mammals. Coalescent-based analyses of more than 600-Mbp genomic data and network analyses show that the genomes of rorquals are characterized by contradicting genealogies for their central divergence. Thus, the evolution of rorquals appears to be a process of gradual divergences that likely gave rise to three lineages almost simultaneously: (i) blue plus sei whales, (ii) gray

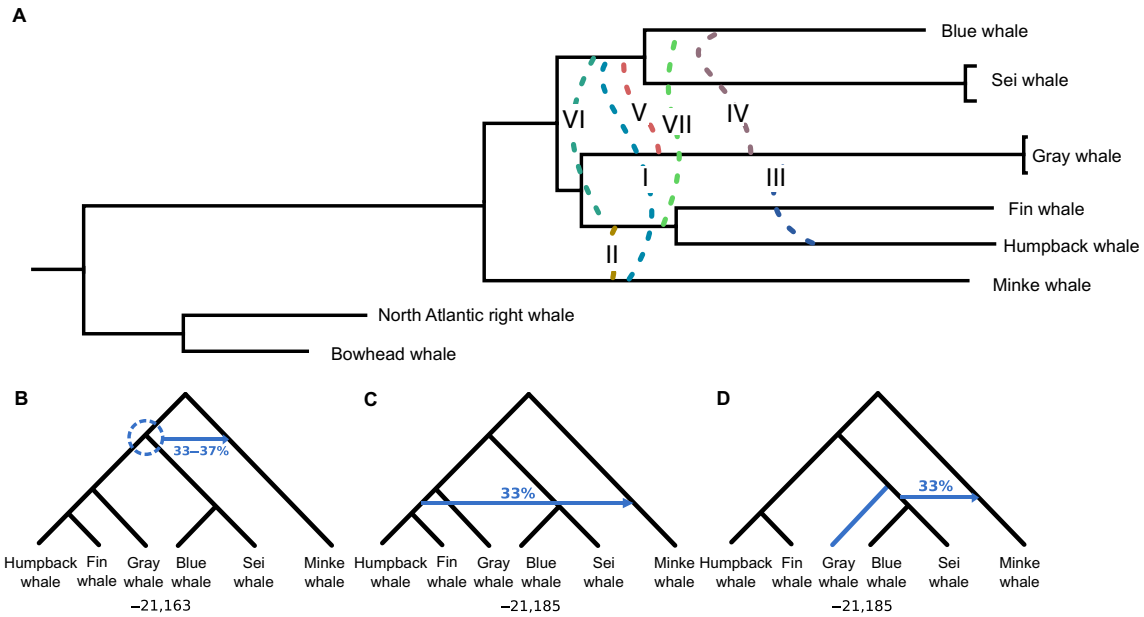


Fig. 3. Gene flow signals for baleen whales inferred by the *D* statistic, D_{FOIL}, and PhyloNet. (**A**) The species tree of baleen whales with gene flow signals detected by the *D* statistic and D_{FOIL} indicated by dashed lines. Signals I to IV were inferred by the *D* statistic, and signals V, VI, and VII were detected by D_{FOIL} and were partially corroborated by the *D* statistic. Note that D_{FOIL} cannot infer gene flow involving the minke whale. (**B** to **D**) Rooted networks for the Balaenopteridae sensu lato phylogeny with reticulations inferred from PhyloNet based on 34,192 20-kbp GFs. Reticulations are shown as blue arrows with inheritance probability denoted above or below. Log-likelihood scores are shown below the networks. Notably, inheritance probability around 33% resembles the distribution of quartet scores and the phylogenetic signals from GFs (Fig. 1). (B) The three best networks indicated a reticulation originating at the circled three branches to minke whale. Similar likelihood scores do not allow the identification of a single origin of gene flow; therefore, the networks were merged, and a range of inheritance probabilities is given. (C) The fourth best network has only a marginally poorer likelihood score and indicates a reticulation between the ancestor of the fin and humpback whale and that of the minke whale. (D) The fifth best network has the same likelihood as (C) and finds an alternative placement of gray whale (blue branch) and reticulation from the ancestor of the blue and sei whale to that of the minke whale.

Downloaded from http://advances.sciencemag.org/ on April 18, 2018

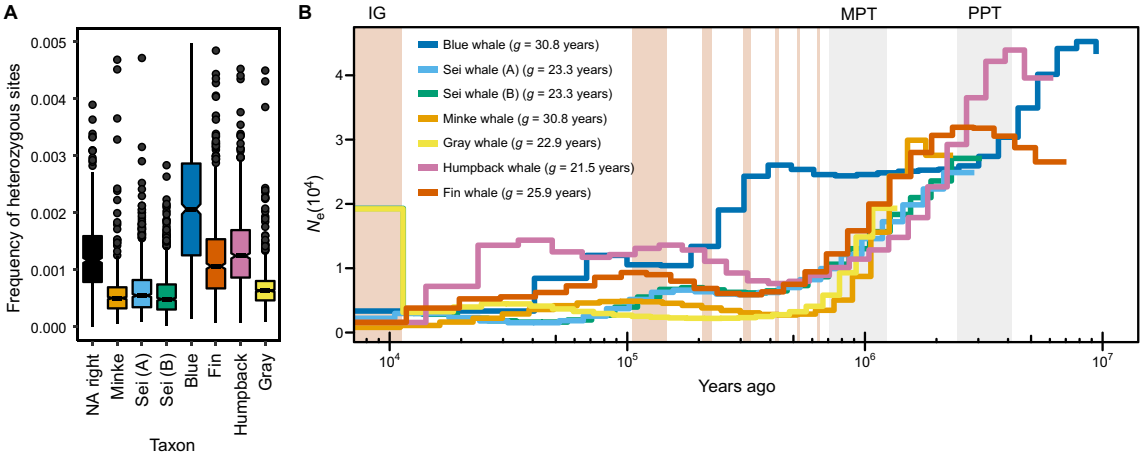


Fig. 4. Demographic history and genome-wide heterozygosity. (**A**) Genome-wide heterozygosity estimated from genomic 100-kbp windows. (**B**) Historical N_e using the PSMC analyses for all baleen whale genomes. The *x* axis shows the time, and the *y* axis shows N_e. Plots were scaled using a mutation rate (μ) of 1.39×10^{-8} substitutions nucleotide^{-1} generation^{-1} and species-specific generation times (*g*). Generation times are noted next to the species names. Light brown shading indicates interglacials (IG) in the Pleistocene and Holocene, and gray shading indicates the MPT and the PPT.

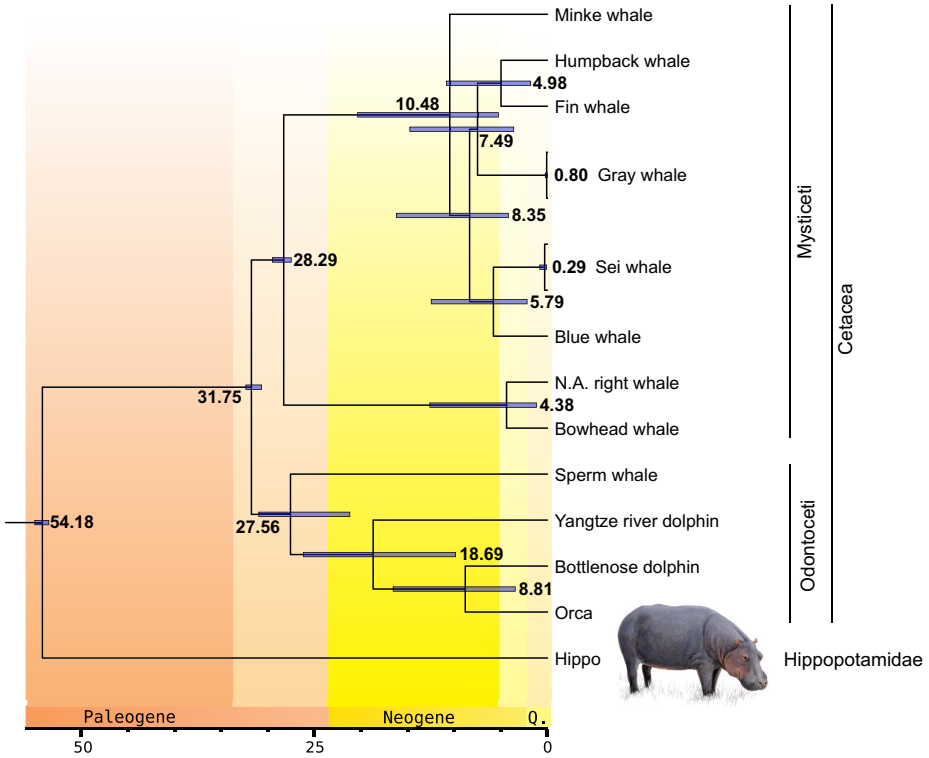


Fig. 5. Divergence time tree of Cetancodonta (*56*) including the newly sequenced baleen whales, estimated from 234,947 amino acid sites (2778 orthologs). Rorquals diverged in the late Miocene, 10.5 to 7.5 Ma ago. Four other cetartiodactyl species were also included but not shown due to space constraints; the dog (*Canis lupus familiaris*) was used as an outgroup. Five calibration points were used for dating (table S8) (*29*, *56*–*60*).

Downloaded from http://advances.sciencemag.org/ on April 18, 2018

whale, and (iii) fin plus humpback whales. The early rorqual radiation is therefore best understood as a phylogenetic network because different fragments of the rorqual genomes support three different evolutionary histories. This provides the reason why the evolution of rorquals was previously differently reconstructed and poorly supported by molecular analyses of smaller data sets (*5*–*8*). Their evolutionary reconstruction needed to be constrained by morphological data to yield a traditional bifurcating tree among rorquals (*2*).

The apparently unequivocal support for the species tree by the MSC analyses is likely a consequence of a slight imbalance of the evolutionary signal that preferably places the gray whale together with the fin whale and humpback whale. Within the massive amount of genome-scale data, even a minor bias can lead to significantly resolved branches, despite the underlying conflict (*31*). Therefore, inspection of quartet scores in a coalescent species tree and network and CONSENSE analyses are crucial in identifying and depicting conflict in the evolutionary signal.

Rorqual taxonomy

Despite the conflict for the early divergence among rorquals, other divergences are well resolved by genome analyses that find the humpback whale closely related to the fin whale within the genus *Balaenoptera*. This is consistent with previous mitogenomic studies (*5*, *7*, *21*) and makes a separate genus, *Megaptera*, obsolete. If the rules of scientific nomenclature are strictly followed in accord with the phylogenetic relationships, the preferred name of the humpback whale should be *Balaenoptera novaeangliae*.

Because gray whales are morphologically, behaviorally, and ecologically distinct from other balaenopterid whales, placing them in a separate family (Eschrichtiidae) distinct from Balaenopteridae sensu stricto seemed natural (*1*, *32*). This classification has been questioned by some molecular analyses (*5*, *21*), and the current genomic analyses resolve this issue conclusively. Despite their derived morphology, gray whales fall unquestionably within the genus *Balaenoptera*, challenging their status as a separate family or even as a separate genus. Notably, the first described specimen of a gray whale was named *Balaenoptera robusta* (*33*) but later classified as own family and genus by J.E. Gray in 1865 in honor of the zoologist D. F. Eschricht (*32*). Consequently, we suggest that the originally proposed scientific name of the gray whale should be resurrected, with its name included in the Balaenopteridae.

Mechanisms of the rorqual radiation

The radiation of extant rorquals is documented by a rich fossil record with a notable diversity of evolutionary distinct lineages, most of which are now extinct. Speciation is generally assumed to occur when biological or geographic isolation results in reproductive isolation (*34*), and it may be difficult to conceive how whales could diverge. Compared to the terrestrial environment, the marine realm is a three-dimensional continuum, almost devoid of barriers that could aid allopatric speciation for highly mobile organisms such as whales. Mixing of gene pools among rorquals can still occur, and such a process would hinder diversification and consequently speciation (*9*). Even some 8 Ma (or about 400,000 generations ago) after their initial divergence, some baleen whale species can still hybridize, which might also be facilitated by their strikingly uniform karyotypes (*11*).

However, ongoing sympatric speciation in marine mammals by the formation of discrete ecotypes has been suggested for the orca or killer whale (*Orcinus orca*) (*35*). For example, the so-called "transient" and "resident" ecotypes specialized to prey on mammals and fish, respectively (*35*). Similarly, rorquals have evolved different feeding strategies. Whereas most baleen whales feed on pelagic prey such as zooplankton and small fish, the gray whales have evolved to feed on benthic invertebrates by scooping up the seafloor. This opened a new ecological niche to which the gray whale adapted, leading over time to sympatric speciation. The adaptation to the benthic food source also led to notable morphological changes, consequently placing the gray whale into an own family. This differentiation may be triggered by climatic change and other environmental disturbances. These different ecological specializations could have led to a speciation continuum in the past that is similar to the one observed in orcas today.

Genomic analyses find the divergence times of baleen whales to be somewhat younger but within the range of previous estimates (*5*, *8*, *21*). The rorqual radiation coincides with the late Miocene cooling at ~7 Ma ago (*36*). This global cooling affected the marine environment by the onset of the current equator-pole temperature gradient. The beginning of the modern oceanic circulation increased productivity in the temperate and polar oceans (*36*), which may have affected cetacean evolution into different ecotypes.

Network-like evolution in whales

It seems counterintuitive that even whole-genome data do not fully resolve the evolution of whales and other mammals in a bifurcating pattern (*12*). However, speciation being a continuous process with possible hybridization, rather than a strict dichotomous event, has already been recognized by Darwin (*37*) and has recently gained new attention (*38*). In sympatric speciation, genomes can be homogenized by gene flow, and only a few genes need to be under divergent selection to form new species (*38*). Genome analyses sometimes fail to support the idea that speciation by reproductive isolation can fail to yield a fully resolved bifurcating tree, which has been the ultimate goal of evolutionary studies for many years. The analysis of genome sequences rather allows observing and comprehending evolutionary incongruence to translate this into new evolutionary hypotheses that might be better depicted as networks (*39*). Recognizing that "divergence with genetic exchange" is a widespread phenomenon in animals (*9*) makes it necessary to review the biological species concept. Instead of relying on reproductive isolation (*34*), a modern species concept should incorporate selective processes that maintain species divergence even under gene flow (*12*).

Signals for introgressive hybridization

Signals for gene flow confirm sightings and reports of current hybridization in whales (*10*, *40*, *41*). The signal for gene flow between blue and fin whale confirms introgression in these species. Other reports on hybrids between humpback and blue whales (*40*) or between bowhead and right whales (*42*) could not be confirmed by the present genome analyses. The hybridization between these species is likely restricted to few individuals or populations and did not lead to introgression. Further sequencing efforts will give more detailed insights into the extent of introgression of baleen whales and potential ecological implications.

In recent genomic studies of bears, humans, and many other animals, gene flow from introgressive hybridization has been identified as a cause for phylogenetic incongruence (*9*, *12*). Postspeciation gene flow can be analyzed in genomic data with a variety of methods (*43*). The *D* statistic and its derivative are undoubtedly the widest applied methodology (*24*, *25*), but these approaches assume a fully resolved species tree. If the species tree includes polytomies or, based on inappropriate statistical methods, is misidentified (*44*), then the basic assumption of the *D* statistic may be violated and the results can be misleading. Therefore, in case of phylogenetic uncertainties, gene flow

Downloaded from http://advances.sciencemag.org/ on April 18, 2018

analyses should, in addition, apply methods that do not require a known topology such as PhyloNet that infers introgression signals from a set of gene trees (*26*). However, alternative methods can be computationally intractable for complex phylogenies or a large number of loci.

Demographic history

Genome data from a single individual allow the reconstruction of the effective population size of its species for some 1 to 2 Ma back in time (*28*). These studies have shown that the demographic histories of many mammals have been influenced by climatic oscillations in the Pleistocene [for example, sheep (*45*)]. However, baleen whales maintained relatively stable effective population sizes after the MPT, despite major oscillations in the global climate consequently affecting ocean circulation, upwelling, and marine productivity. The general congruence of population size histories of different baleen whale species indicates that they were similarly affected by these factors. Differences in sequence depth may limit the comparison of absolute N_e between our samples; however, chronology of the curves is not expected to be affected (*46*). Industrial whaling has been too recent to leave a noticeable signal of a declining N_e in the PSMC analyses, especially for long-lived species with long generation times like rorquals. However, compared to other mammals, rorquals, particularly the blue whale, have a comparatively high degree of genome-wide heterozygosity (*27*). The impact of whaling on the genetic diversity of baleen whales may become apparent only after several generations and require population-scale studies for a detailed assessment (*47*).

CONCLUSION

Genome data analyses finally resolved the evolutionary history of baleen whales, even if it is not a bifurcating tree that most had expected. The evolution of rorquals can only be accurately understood by phylogenetic networks because a forced bifurcating tree or a hard polytomy would ignore the accumulated evolutionary history that is recorded in their genomes. It is evident that the central rorqual radiation was not along a progressively ordered process. On the contrary, speciation with gene flow is indicated by the nearly equal probabilities for different evolutionary histories across rorqual genomes. In addition, hybridization between blue and fin whales left genome-wide signals of introgression. The gray whale may constitute a striking example of sympatric speciation related to adaptation to and occupation of a particular niche, bottom feeding, as compared to the pelagic feeding of other rorquals. Our results indicate that sympatric speciation should not be neglected as a mode of speciation in highly connected habitats, such as the marine environment.

MATERIALS AND METHODS

DNA isolation and sequencing

Cell cultures (established by the first author, 1969 to 1974) were grown in Dulbecco's modified Eagle's medium supplemented with 10% fetal bovine serum under standard conditions. DNA of *H. amphibius* was extracted from muscle tissue of a naturally deceased individual, provided by M. Bertelsen (Copenhagen Zoo). DNA was isolated from cells or tissue using a standard phenol-chloroform method. Sequencing libraries were prepared with insert sizes between 300 and 500 bp and sequenced using Illumina HiSeq 2000, 2500, and 4000 technology. The minke whale genome data were obtained from the short read archive (accession no. SRR896642) (*15*). Sequencing library information and mapping statistics are given in table S1. Quality control was performed using FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc/), and reads were trimmed. All cell culture work and DNA extractions from tissues were performed according to the ethical guidelines and permission of the respective institutions.

Paired-end reads were mapped to the bowhead whale genome (*B. mysticetus*) (*16*), with BWA mem version 0.7.12-r1039 (*48*), and duplicates were marked with picard (https://github.com/broadinstitute/picard). The bowhead whale was used as reference genome because it avoids a mapping bias that can affect phylogenetic analyses. The minke whale is phylogenetically placed inside baleen whales, and a possible mapping bias against its genome is likely to affect phylogenetic and gene-flow analyses. Scaffolds shorter than 100 kbp were excluded. Repetitive sequences were annotated for the bowhead whale genome by RepeatMasker (*17*). From the mapped reads, single-nucleotide variants (SNVs) and short insertion or deletions (InDels) were called by freebayes v0.9.20-16-g3e35e72 (*49*) with a minimum coverage of four reads and settings: –-monomorphic –min-mapping-quality 20, –C 4, –F 0.3. Consensus sequences were created from VCF-files using custom perl scripts. InDels were removed, and ambiguously called sites were masked as "N."

For sequencing the hippopotamus genome, paired-end and mate-pair libraries were constructed with different insert sizes sequenced on Illumina HiSeq 2000/2500 sequencers (table S4). Because of high levels of duplications, mate-pair libraries were deduplicated. All libraries were trimmed for adaptors and low-quality regions, requiring a minimum read length of 90 bp after trimming. All libraries were assembled into contigs using Minia with $k = 49$ (*19*). Contigs were scaffolded with SSPACE (https://github.com/nsoranzo/sspace_basic) using the mate-pair libraries. Finally, GapCloser (http://soap.genomics.org.cn/) was run with all libraries. Scaffolds shorter than 1 kbp were excluded from the final genome assembly of the hippopotamus. Novel repetitive elements were identified with RepeatModeler (www.repeatmasker.org/RepeatModeler/).

The genome assembly was screened for repetitive sequences using RepeatMasker and the previously created de novo library of identified repeats from RepeatModeler and the RepBase Mammalia library. To account for nonoverlapping detected repeats, we combined and applied the genome masks to the genome sequence. Protein coding genes were predicted ab initio with AUGUSTUS v.3.1 (*20*) using settings –UTR -species human.

Phylogenomic analysis of baleen whales

Consensus sequences of all genomes were aligned per scaffold, and heterozygous sites and repetitive regions were removed. Per-scaffold alignments were split into nonoverlapping GFs of 10, 20, and 100 kbp, respectively. Scaffolds that were shorter than the GF size after removal of ambiguous sites were excluded.

Estimating phylogenetic information in GFs

To analyze the phylogenetic information content of the GFs, we randomly sampled 5000 GFs to count the number of parsimony informative sites and to estimate the genetic distance between the two closest related whales, that is, the bowhead and the North Atlantic right whale. On the basis of real GFs, we simulated GFs between lengths of 1 and 100 kbp to determine which length carries sufficient phylogenetic information to statistically reject alternative topologies (fig. S1). Topology testing was performed using the AU test (*18*).

Species-tree inference and analysis of phylogenetic conflict

JModelTest2 (*50*) identified the suitable nucleotide substitution model by evaluating random 20-kbp GFs. For each GF, phylogenetic trees were

Downloaded from http://advances.sciencemag.org/ on April 18, 2018

computed with RaxML (*51*) using ML and the GTR + G substitution model that was identified as best fit. Each ML analysis was bootstrapped with 100 replicates. From all 20-kbp GF trees, ASTRAL 4.10.5 (*31*) computed a species tree under the MSC model (exact method) returning quartet scores and posterior probabilities. The species tree was rooted with the bowhead whale and North Atlantic right whale that are outside Balaenopteridae. CONSENSE from the PHYLIP package (*22*) explored conflict among the gene trees by identifying identical splits in a set of given gene trees and summarizing their frequency. Consensus networks of the GF trees were generated using SplitsTree4 (*23*) with different median thresholds. Phylogenetic consensus networks summarize gene tree discordance by drawing alternative edges for each observed split.

Phylogeny of whale mitochondrial genomes

We reconstructed the mitochondrial (mt) genomes from the whale individuals reported herein by mapping the reads to conspecific published mt genomes and generated consensus sequences as described for the nuclear genomes. Mt sequences were aligned to 19 published mt sequences of whales. Accession numbers of mt genomes used as reference for mapping and the phylogenetic analysis are shown in fig. S4. A Bayesian phylogenetic tree was reconstructed using MrBayes version 3.2.2. The analysis was run for 1,200,000 generations with default priors, using the "invgamma" substitution model and an arbitrary burn in of 25% of the samples.

Gene flow analyses

The *D* statistic compares the number of biallelic ABBA and BABA sites in a four-taxon phylogeny and requires a phylogenetic topology following (((H1, H2), H3), O), with H1 to H3 being ingroups and O being the outgroup. For the analyses, the consensus sequences of baleen whales were fragmented into nonoverlapping 100-kbp windows. We applied the *D* statistic to all asymmetric four-taxon phylogenies that can be extracted from the species tree. This resulted in 33 gene flow analyses, such as "(((blue whale, sei whale), fin whale), minke whale)." The direction of gene flow can be estimated in a derivative of the *D* statistic, the D_{FOIL} analysis (*25*), downloaded 15 September 2015 from https://github.com/jbpease/dfoil. The test requires an asymmetric five-taxon tree with a specific topology; therefore, not all combinations of five whale taxa could be analyzed. The D_{FOIL} analyses used the same genomic windows as the *D* statistic analyses.

Our taxon sampling allowed the analysis of the following topologies when considering the estimated species tree as correct because the D_{FOIL} analyses assume a symmetrical five-taxon topology: (i) (((blue, sei), (fin, hump)), NA right); (ii) (((blue, sei), (fin, gray)), minke); (iii) (((blue, sei), (hump, gray)), minke); (iv) (((blue, sei), (hump, gray)), NA right); (v) (((blue, sei), (hump, gray)), bowhead); (vi) (((blue, sei), (fin, gray)), NA right); (vii) (((blue, sei), (fin, gray)), bowhead); (viii) (((blue, sei), (fin, hump)), bowhead); NA right refers to the North Atlantic right whale, whereas the remaining whales are indicated by the first part of their common names.

Maximum likelihood inference for reticulation with PhyloNet

PhyloNet (*26*) is specifically developed to reconstruct reticulated phylogenies from a set of gene trees. We used the ML approach to analyze a set of every 10th GF ML tree, that is, 3419 trees in a coalescent framework that accounts for ILS while allowing different numbers of reticulations (*26*). Subsampling of trees reduced complexity and computational demand. In addition, the bowhead whale, North Atlantic right whale, and sei whale individual "B" were pruned from the input gene trees because their phylogenetic position is unambiguous. The "InferNetwork_ML" method was run with 50 iterations, yielding the five networks with the highest likelihood scores. Analyzing networks with more than one reticulation were too complex and not interpretable from the extended Newick format.

Demographic history

Changes in N_e for the baleen whales were inferred from genome sequences using the PSMC (*28*). We applied PSMC v0.6.5-r67 with input files generated using Samtools mpileup version 1.2 (www.htslib.org) and by applying a minimum mapping and base quality of 30. Using vcfutils, minimum and maximum depth of coverage thresholds were set to 0.5 and 2× the sample's average coverage (table S1). PSMC was run with 25 iterations, an N_0-scaled maximum coalescent time of 20, and a ρ/θ ratio of 5, and the 64 time intervals were parameterized as "4 + 25 × 2 + 4 + 6." PSMC plots were scaled with a mutation rate of $\mu = 4.5 \times 10^{-10}$ mutations bp^{-1} $year^{-1}$ that has been determined for whales (*52*).

Bootstrapping was performed on whole scaffolds. Species-specific predisturbance generation times were used to scale the PSMC plots (*53*). Industrial whaling took place only during the last 200 years, so predisturbance generation times are more accurate for the time frame covered by PSMC. The generation times are shown in Fig. 5.

Genome-wide heterozygosity

To estimate the genome-wide heterozygosity, we randomly sampled 1000 100-kbp nonoverlapping windows for each genome. For these windows, heterozygous SNVs were extracted from the complete set of called variants. Heterozygous sites were excluded if the distance to a called InDel was 10 bp or less or if the sequencing depth at the site was less than 0.5 or 2× the mean sample coverage. This avoids artifacts from assembly errors. For each window, the frequency of heterozygous sites was calculated. In addition, genome-wide heterozygosity and genome-wide sequencing error were inferred using mlRho (*54*). To exclude the potential effects of higher sequencing coverage in the blue whale, the BAM file was downsampled using GATK (genome analysis tool kit) and genome-wide heterozygosity was estimated for ~10× sequencing data.

Cetartiodactyla phylogenomics

Protein sequences for different representative species among Cetartiodactyla were retrieved from ENSEMBL and RefSeq (table S7). For data obtained from RefSeq, Samtools extracted the CDSs from whole-genome sequences using the annotation provided as a General Feature Format (GFF) file.

The annotated CDS for the bowhead whale was used to extract and translate the corresponding genomic regions from baleen whale genomes that were mapped to the bowhead whale Proteinortho version 5.11 screened protein sequences from all genomes listed in table S7. The baleen whale genomes were mapped to the bowhead whale genome and thus their CDSs have the same genomic coordinates. Therefore, the protein sequences of the baleen whales were added after orthology detection based on orthologous proteins identified in the bowhead whale. All proteins for which orthologs were identified in at least nine species were selected, and their sequences were extracted. Protein sequences were aligned individually and trimmed to exclude ambiguously aligned sites. The trimmed alignments were concatenated and used to date the

Downloaded from http://advances.sciencemag.org/ on April 18, 2018

cetartiodactyl species tree with MCMCTree (*55*) using five calibration points across the tree of Cetartiodactyla (table S8).

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/4/4/eaap9873/DC1

fig. S1. Possible tree topologies for baleen whales that were evaluated by the AU test.
fig. S2. Phylogenetic content of GFs.
fig. S3. AU test for increasing GF sizes.
fig. S4. MSC-based species trees generated by ASTRAL using 34,192 GFs, with each GF being 20 kbp long.
fig. S5. Phylogenetic tree from mitochondrial genomes for baleen whales.
fig. S6. A majority-rule consensus tree from 34,192 individual GF ML trees (table S6) calculated with the program CONSENSE of the PHYLIP package.
fig. S7. Consensus networks for baleen whales from 34,192 gene trees (10-kbp GF) at different minimum thresholds of gene trees to form an edge.
fig. S8. ML estimates of genome-wide heterozygosity estimated with mlRho.
fig. S9. Blue whale heterozygosity for different sequencing depth.
fig. S10. Demographic histories for each individual whale genome with 100 bootstrap replicates.
table S1. Sequencing and mapping statistics.
table S2. Occurrences of repetitive elements in the bowhead whale genome.
table S3. Number of called substitutions for each whale genome.
table S4. Library and sequencing information for the hippopotamus genome assembly.
table S5. Summary of repetitive elements in the hippopotamus genome.
table S6. A majority-rule consensus analysis of 34,192 individual GF ML trees.
table S7. Common names, scientific names, accession numbers, and source database of additional genomes that were included in the divergence time analyses.
table S8. Calibration points used for the divergence time tree, node age estimates in million years ago, and references.
table S9. Divergence time estimates for Artiodactyla and Cetacea for nodes in the divergence time tree (Fig. 5).
data S1. *D* statistics results.
data S2. D_{FOIL} results.

REFERENCES AND NOTES

1. R. M. Nowak, *Walker's Mammals of the World* (Johns Hopkins Univ. Press, ed. 6, 1999).
2. F. G. Marx, R. E. Fordyce, Baleen boom and bust: A synthesis of mysticete phylogeny, diversity and disparity. *R. Soc. Open Sci.* **2**, 140434 (2015).
3. A. Werth, in *Feeding: Form, Function, and Evolution in Tetrapod Vertebrates*, K. Schwenk, Ed. (Academic Press, 2000), pp. 487–526.
4. G. J. Slater, J. A. Goldbogen, N. D. Pyenson, Independent evolution of baleen whale gigantism linked to Plio-Pleistocene ocean dynamics. *Proc. Biol. Sci.* **284**, 20170546 (2017).
5. A. Hassanin, F. Delsuc, A. Ropiquet, C. Hammer, B. Jansen van Vuuren, C. Matthee, M. Ruiz-Garcia, F. Catzeflis, V. Areskoug, T. T. Nguyen, A. Couloux, Pattern and timing of diversification of Cetartiodactyla (Mammalia, Laurasiatheria), as revealed by a comprehensive analysis of mitochondrial genomes. *C. R. Biol.* **335**, 32–50 (2012).
6. M. Nikaido, H. Hamilton, H. Makino, T. Sasaki, K. Takahashi, M. Goto, N. Kanda, L. A. Pastene, N. Okada, Baleen whale phylogeny and a past extensive radiation event revealed by SINE insertion analysis. *Mol. Biol. Evol.* **23**, 866–873 (2006).
7. T. Sasaki, M. Nikaido, H. Hamilton, M. Goto, H. Kato, N. Kanda, L. Pastene, Y. Cao, R. Fordyce, M. Hasegawa, N. Okada, Mitochondrial phylogenetics and evolution of mysticete whales. *Syst. Biol.* **54**, 77–90 (2005).
8. U. Arnason, A. Gullberg, A. Janke, Mitogenomic analyses provide new insights into cetacean origin and evolution. *Gene* **333**, 27–34 (2004).
9. M. L. Arnold, *Divergence with Genetic Exchange* (Oxford Univ. Press, 2015).
10. R. Spilliaert, G. Vikingsson, U. Arnason, A. Palsdottir, J. Sigurjonsson, A. Arnason, Species hybridization between a female blue whale (*Balaenoptera musculus*) and a male fin whale (*B. physalus*): Molecular and morphological documentation. *J. Hered.* **82**, 269–274 (1991).
11. U. Árnason, I. F. Purdom, K. W. Jones, Conservation and chromosomal localization of DNA satellites in balenopterid whales. *Chromosoma* **66**, 141–159 (1978).
12. V. Kumar, F. Lammers, T. Bidon, M. Pfenninger, L. Kolter, M. A. Nilsson, A. Janke, The evolutionary history of bears is characterized by gene flow across species. *Sci. Rep.* **7**, 46487 (2017).
13. B. M. Hallström, A. Janke, Mammalian evolution may not be strictly bifurcating. *Mol. Biol. Evol.* **27**, 2804–2816 (2010).
14. J. H. Degnan, N. A. Rosenberg, Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* **24**, 332–340 (2009).
15. H.-S. Yim, Y. S. Cho, X. Guang, S. G. Kang, J.-Y. Jeong, S. S. Cha, H.-M. Oh, J.-H. Lee, E. C. Yang, K. K. Kwon, Y. J. Kim, T. W. Kim, W. Kim, J. H. Jeon, S.-J. Kim, D. H. Choi, S. Jho, H.-M. Kim, J. Ko, H. Kim, Y.-A. Shin, H.-J. Jung, Y. Zheng, Z. Wang, Y. Chen, M. Chen, A. Jiang, E. Li, S. Zhang, H. Hou, T. H. Kim, L. Yu, S. Liu, K. Ahn, J. Cooper, S.-G. Park, C. P. Hong, W. Jin, H.-S. Kim, C. Park, K. Lee, S. Chun, P. A. Morin, S. J. O'Brien, H. Lee, J. Kimura, D. Y. Moon, A. Manica, J. Edwards, B. C. Kim, S. Kim, J. Wang, J. Bhak, H. S. Lee, J.-H. Lee, Minke whale genome and aquatic adaptation in cetaceans. *Nat. Genet.* **46**, 88–92 (2014).
16. M. Keane, J. Semeiks, A. E. Webb, Y. I. Li, V. Quesada, T. Craig, L. B. Madsen, S. van Dam, D. Brawand, P. I. Marques, P. Michalak, L. Kang, J. Bhak, H.-S. Yim, N. V. Grishin, N. H. Nielsen, M. P. Heide-Jørgensen, E. M. Oziolor, C. W. Matson, G. M. Church, G. W. Stuart, J. C. Patton, J. C. George, R. Suydam, K. Larsen, C. López-Otín, M. J. O'Connell, J. W. Bickham, B. Thomsen, J. P. de Magalhães, Insights into the evolution of longevity from the bowhead whale genome. *Cell Rep.* **10**, 112–122 (2015).
17. A. F. A. Smit, R. Hubley, P. Green, RepeatMasker Open-3.0 (2010); www.repeatmasker.org.
18. H. Shimodaira, An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* **51**, 492–508 (2002).
19. R. Chikhi, G. Rizk, Space-efficient and exact de bruijn graph representation based on a bloom filter, in *Algorithms in Bioinformatics*, B. Raphael, J. Tang, Eds. (Springer, 2012), vol. 7534, pp. 236–248.
20. M. Stanke, O. Schöffmann, B. Morgenstern, S. Waack, Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**, 62 (2006).
21. M. R. McGowen, M. Spaulding, J. Gatesy, Divergence date estimation and a comprehensive molecular tree of extant cetaceans. *Mol. Phylogenet. Evol.* **53**, 891–906 (2009).
22. J. Felsenstein, PHYLIP—Phylogeny inference package (Version 3.2). *Cladistics* **5**, 164–166 (1989).
23. D. H. Huson, D. Bryant, Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267 (2006).
24. E. Y. Durand, N. Patterson, D. Reich, M. Slatkin, Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28**, 2239–2252 (2011).
25. J. B. Pease, M. W. Hahn, Detection and polarization of introgression in a five-taxon phylogeny. *Syst. Biol.* **64**, 651–662 (2015).
26. Y. Yu, J. Dong, K. J. Liu, L. Nakhleh, Maximum likelihood inference of reticulate evolutionary histories. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 16448–16453 (2014).
27. E. Palkopoulou, S. Mallick, P. Skoglund, J. Enk, N. Rohland, H. Li, A. Omrak, S. Vartanyan, H. Poinar, A. Götherström, D. Reich, L. Dalén, Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth. *Curr. Biol.* **25**, 1395–1400 (2015).
28. H. Li, R. Durbin, Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
29. S. Bajpai, P. D. Gingerich, A new Eocene archaeocete (Mammalia, Cetacea) from India and the time of origin of whales. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 15464–15468 (1998).
30. Z. Liu, M. Pagani, D. Zinniker, R. Deconto, M. Huber, H. Brinkhuis, S. R. Shah, R. M. Leckie, A. Pearson, Global cooling during the Eocene-Oligocene climate transition. *Science* **323**, 1187–1190 (2009).
31. E. Sayyari, S. Mirarab, Fast coalescent-based computation of local branch support from quartet frequencies. *Mol. Biol. Evol.* **33**, 1654–1668 (2016).
32. J. E. Gray, Notes on the whalebone-whales; with a synopsis of the species. *Ann. Mag. Nat. Hist.* **3**, 344–350 (1864).
33. W. Lilljeborg, *Öfversigt af de inom Skandinavien (Sverige och Norrige) anträffade Hvalartade Däggdjur (Cetacea)* (1860).
34. E. Mayr, *Animal Species and Evolution* (The Belknap Press of Harvard Univ. Press, 1963).
35. A. D. Foote, N. Vijay, M. C. Ávila-Arcos, R. W. Baird, J. W. Durban, M. Fumagalli, R. A. Gibbs, M. B. Hanson, T. S. Korneliussen, M. D. Martin, K. M. Robertson, V. C. Sousa, F. G. Vieira, T. Vinař, P. Wade, K. C. Worley, L. Excoffier, P. A. Morin, M. T. Gilbert, J. B. W. Wolf, Genome-culture coevolution promotes rapid divergence of killer whale ecotypes. *Nat. Commun.* **7**, 11693 (2016).
36. T. D. Herbert, K. T. Lawrence, A. Tzanova, L. C. Peterson, R. Caballero-Gill, C. S. Kelly, Late Miocene global cooling and the rise of modern ecosystems. *Nat. Geosci.* **9**, 843–847 (2016).
37. C. Darwin, *On the Origin of the Species* (John Murray, London, 1859).
38. J. L. Feder, S. P. Egan, P. Nosil, The genomics of speciation-with-gene-flow. *Trends Genet.* **28**, 342–350 (2012).
39. E. Bapteste, L. van Iersel, A. Janke, S. Kelchner, S. Kelk, J. O. McInerney, D. A. Morrison, L. Nakhleh, M. Steel, L. Stougie, J. Whitfield, Networks: Expanding evolutionary thinking. *Trends Genet.* **29**, 439–441 (2013).
40. P. A. Folkens, R. R. Reeves, B. S. Stewart, P. J. Clapham, J. A. Powell, *Guide to Marine Mammals of the World* (National Audubon Society, 2002).

Downloaded from http://advances.sciencemag.org/ on April 18, 2018

Genome sequences of the blue whale and other rorquals reveal signatures for introgressive gene-flow.

July 26, 2017

Supplementary Information

Ulfur Arnason[#,1], Fritjof Lammers[#,2,3], Vikas Kumar[2], Maria A. Nilsson[2], Axel Janke[2,3,§]

[#]Shared first authors. [1]Lund University Hospital, Box 117, 221 00 Lund, Sweden. [2]Senckenberg Biodiversity and Climate Research Centre, Senckenberg Gesellschaft für Naturforschung, Senckenberganlage 25, 60325 Frankfurt am Main, Germany. [3]Goethe University Frankfurt, Institute for Ecology, Evolution & Diversity, Biologicum, Max-von-Laue-Str.13, 60439 Frankfurt am Main, Germany. [§]Corresponding author (axel.janke@senckenberg.de)

List of Figures

List of Tables

Supplementary Figures

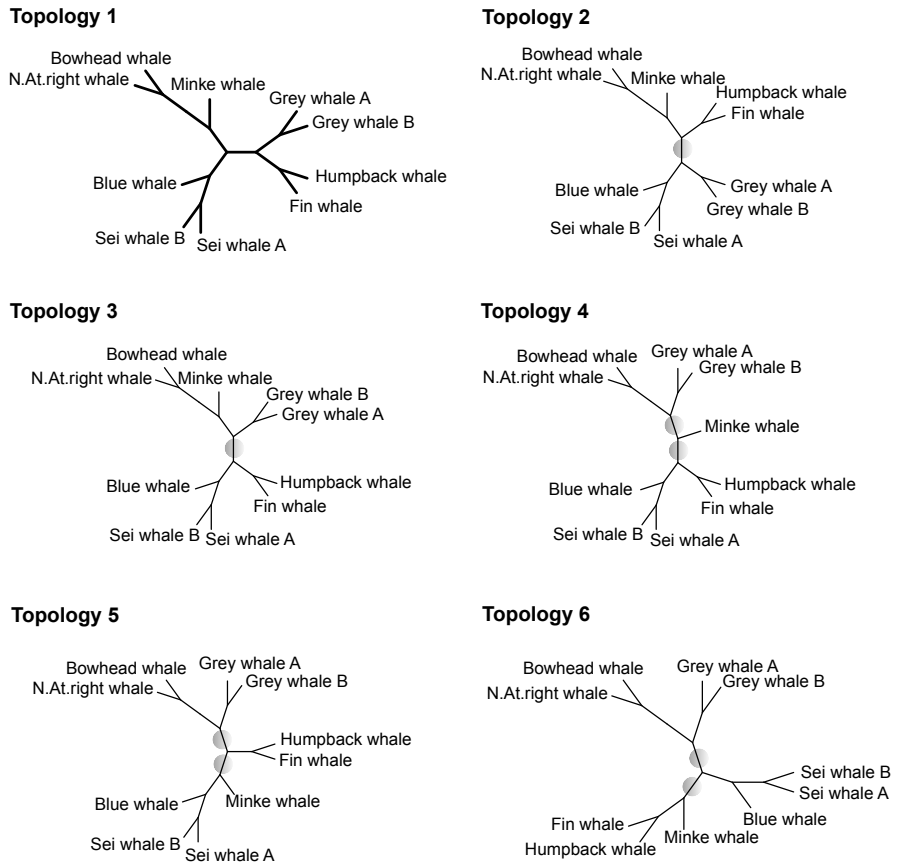


Figure S1. Possible tree topologies for baleen whales that were evaluated by the AU test. Branch swaps that are made relative the species tree (topology 1) are marked by green dots.

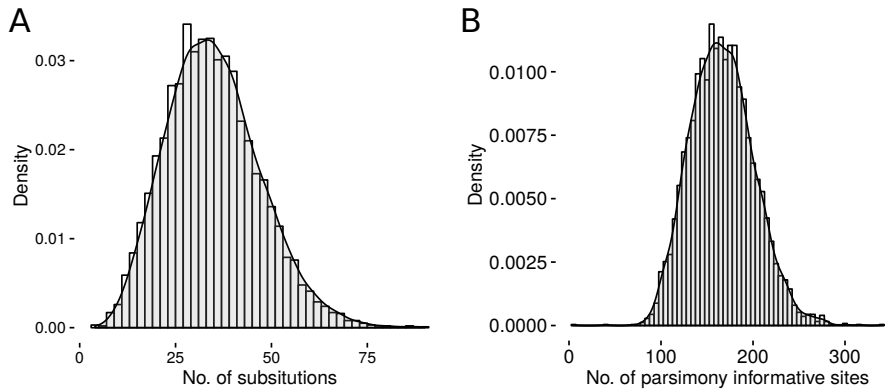


Figure S2. Phylogenetic content of genome fragments. A) Phylogenetic content of 5000 randomly sampled 10 kb genome fragments shown as distribution of absolute genetic distance between the North Atlantic right whale and bowhead whale, two closest related species in the taxon sampling. B) Distribution of parsimony informative sites among 5,000 10kb alignments of the baleen whales.

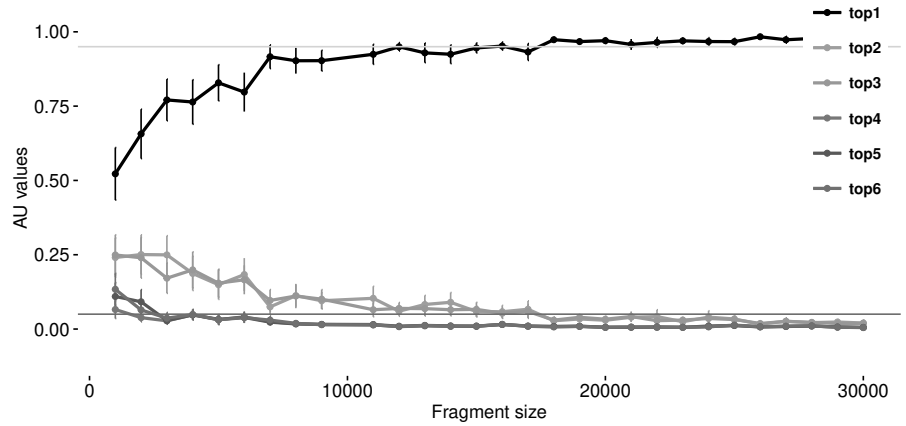


Figure S3. Approximate unbiased test for increasing GF sizes. GF between 1 and 100 kb were simulated using the presumed species tree and real data as input sequence. The AU test evaluates whether the sequence data can statistically support (pAU $>$0.95, green line) or reject (pAU $<$0.05, red line) a phylogenetic hypothesis. The six different topologies are shown in Figure S1.

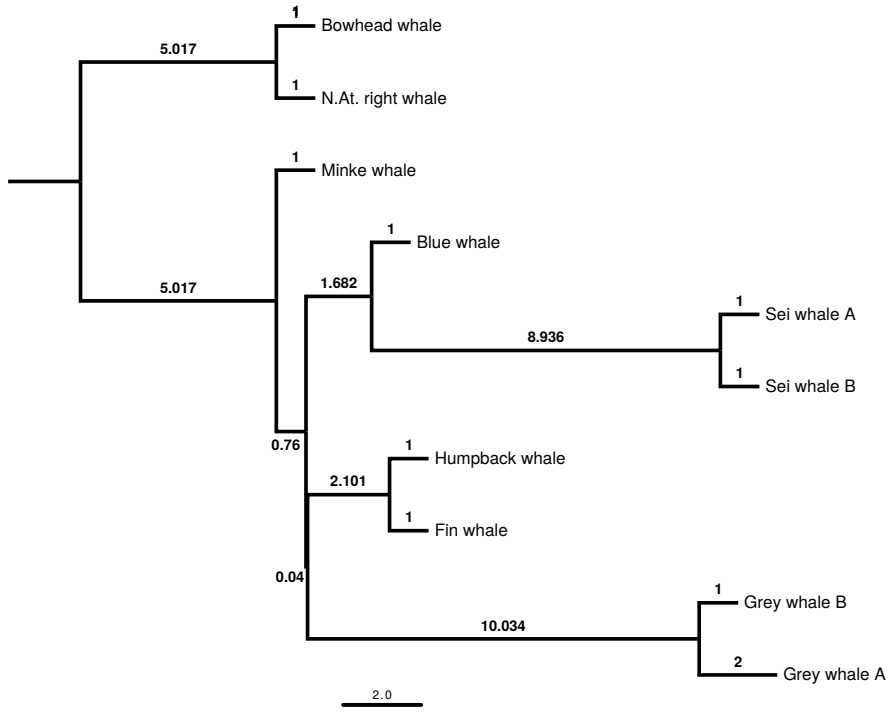


Figure S4. Multi-species coalescent based species tree generated by ASTRAL using 34,192 genome fragments (GF), each 20 kb long. The tree was rooted with Bowhead and North Atlantic right whale. Branch lengths are given in coalescent units and are an indicator of gene-tree discordance, i.e. shorter branches indicate higher gene tree discordance. All branches received unanimous support in the ASTRAL analysis (posterior probability 1.0). N. At. = North Atlantic

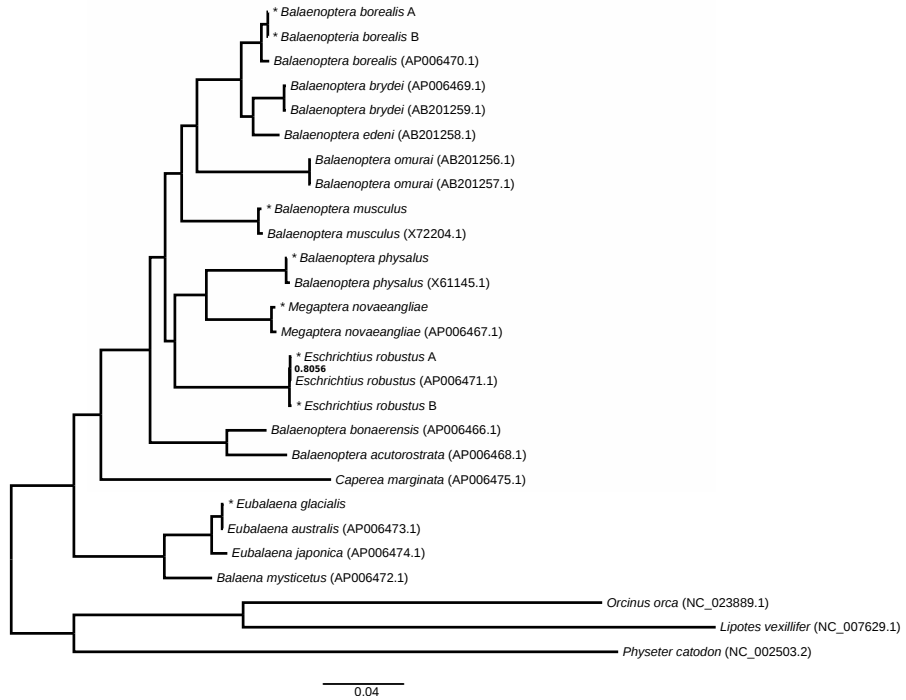


Figure S5. Phylogenetic tree from mitochondrial genomes for baleen whales. New sequences are marked with an asterisk. Accession numbers of published sequences are given in parentheses. Posterior probabilities are given at nodes if not 1.0.

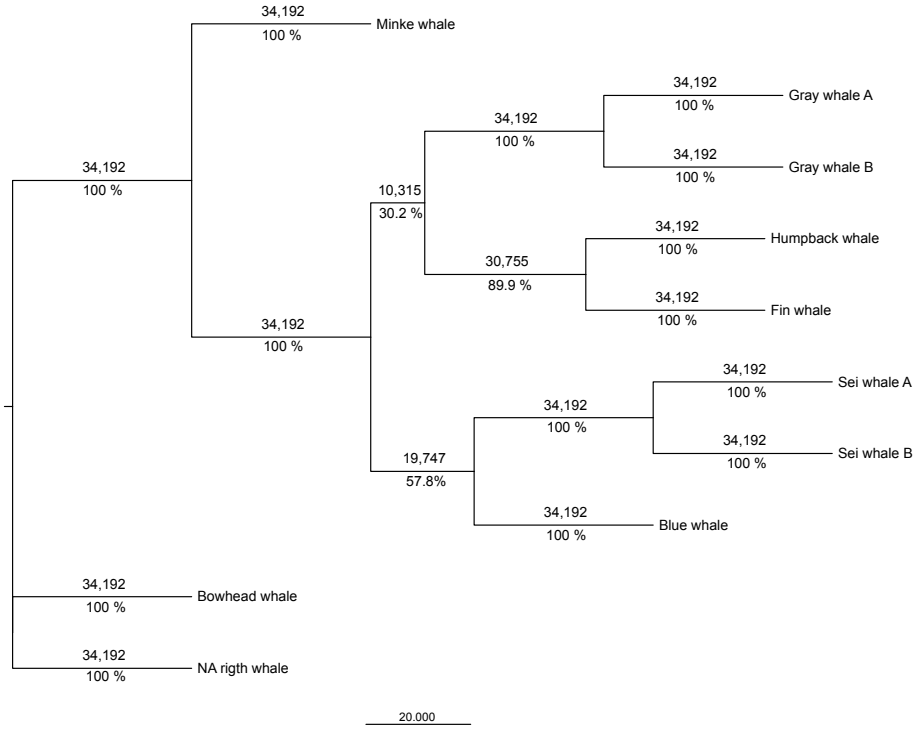


Figure S6. A majority rule consensus tree from 34,192 individual GF ML trees (Table S6) calculated with the program CONSENSE of the PHYLIP package. The topology is congruent to the coalescent species tree. Number above each branches indicate the absolute number of splits found in 34,192 individual GF trees, the number below shows the percentage values.

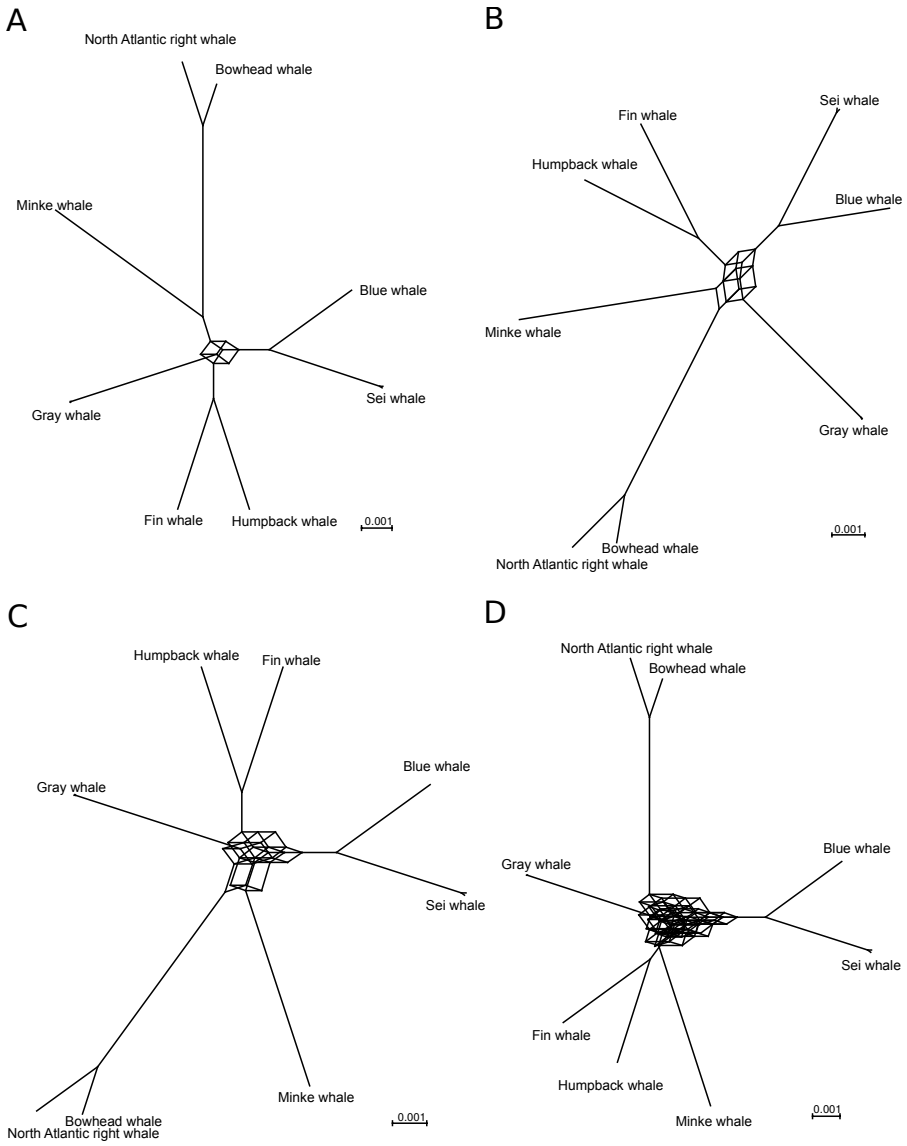


Figure S7. Consensus networks for baleen whales from 34,192 gene trees (10 kb GF) at different minimum thresholds of gene trees to form an edge. A) 14% threshold. B) 11% threshold C) 7% threshold, D) 5% threshold.

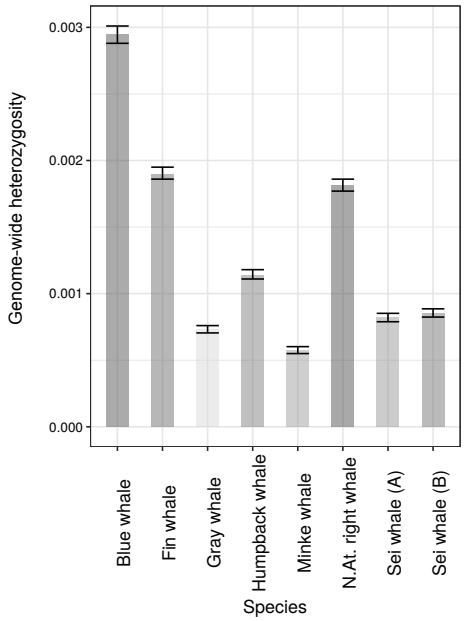


Figure S8. Maximum likelihood estimates of genome-wide heterozygosity estimated with mlRho. Error-bars indicate confidence intervals.

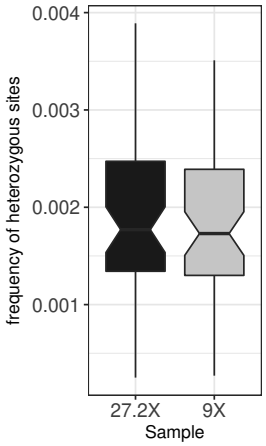


Figure S9. Blue whale heterozygosity for different sequencing depth. The sequencing depth does not affect the frequency of heterozygous sites. Heterozygosity was estimated in 100 kb windows with a total of 5.8 Mb.

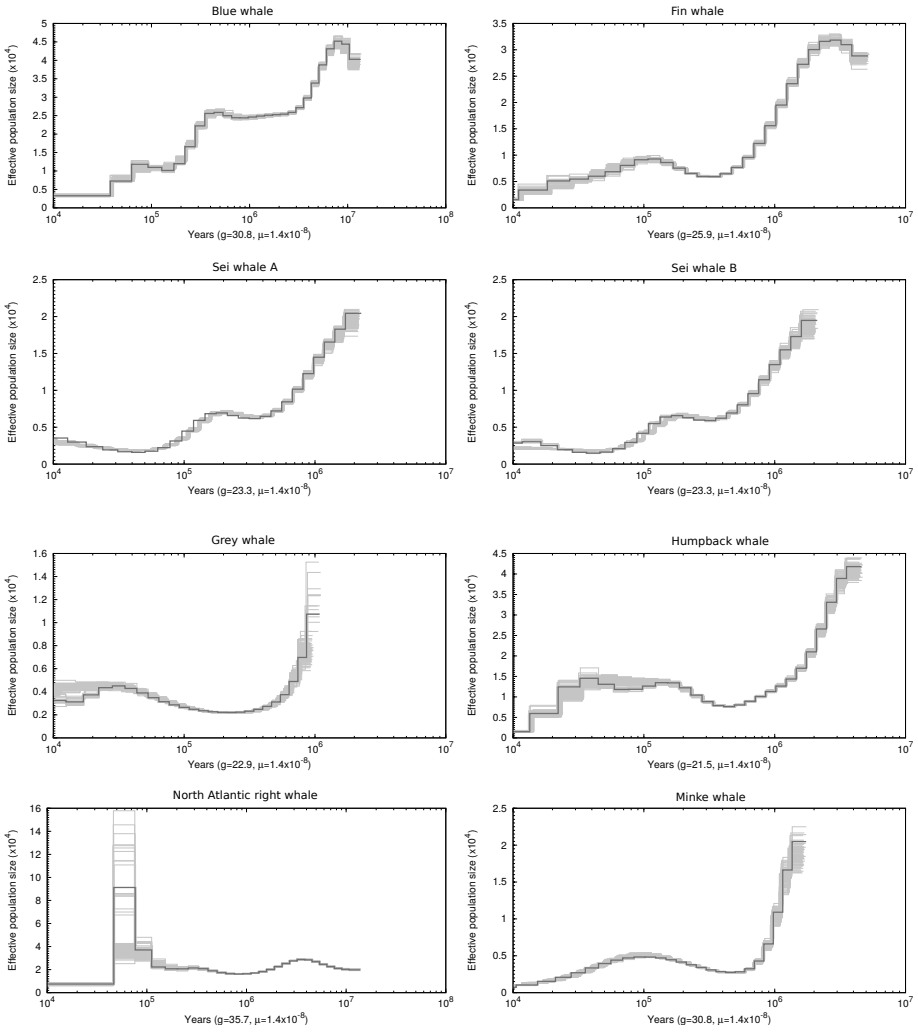


Figure S10. Demographic histories for each individual whale genome with 100 bootstrap replicates. Each panel shows estimated ancestral effective population size calculated by PSMC. Bootstrap replicates are shown in light red. Sample names are given above each panel. Except for North Atlantic right whale, bootstrap replicate indicate little variation in the data. The high degree of variance in bootstrap replicate in North Atlantic right whale likely indicate an artifact in the increase N_e at 10^5 years. g, generation time, μ, mutation rate.

Supplementary Tables

Table S1. Sequencing and mapping statistics. For inidividual whale genome sample, a short ID, common species name, number of generated reads, mean insert size and read length as well asstatistics after mapping to the bowhead whale genome are shown.

ID	Species	No of reads	% mapped	No. duplicates	Coverage	Insert size	Read-length
Egl00	NA right whale	323,959,050	92.3%	43,364,288	8.32x	470	125
Bbo01	Sei whale	316,038,498	91.4%	55,729,034	7.45x	482	125
Bbo02	Sei whale	316,228,639	91.3%	69,502,856	6.93x	482	125
Bmu00	Blue whale	1,131,873,174	92.9%	106,828,877	27.2x	293	100
Mno00	Humpback whale	549,015,152	91.2%	81,954,390	17.0x	470	125
Bph03	Fin whale	415,640,907	93.2%	58,501,974	10.7x	294	100
Bac00	Minke hale	248,445,825	91.9%	63,832,762	7.19x	472	100
Ero01	Gray whale	277,126,828	93.5%	52,844,673	6.30x	330	150
Ero02	Gray whale	349,764,342	93.5%	136,428,269	13.1x	323	150

Table S2. Occurrences of repetitive elements in the bowhead whale genome. Repeats are grouped by their respective repeat family. The table shows the number (count), combined lengths und percentage of the genome sequence.

Family	Count	Length (bp)	Genome-%
SINEs	833,543	140,689,989	6.08
LINEs	1,069,602	426,471,811	18.4
LTR	381,792	139,382,337	6.02
DNA	388,951	83,214,371	3.59
Unclassified	7,269	1,327,012	0.06
Small RNA	293,226	65,978,280	2.85
Satellites	118,187	51,524,590	2.23
Simple repeats	503,891	20,650,898	0.89
Low complexity	84,108	4,216,662	0.18

Table S3. Number of called substitutions for each whale genome. Fixed sites are the number of nucleotide differences compared to the bowhead whale genome sequence.

Sample	Het. sites	Fixed sites
NA right whale	2,675,248	8,538,671
Sei whale A	1,374,610	27,180,101
Sei whale B	1,256,248	26,394,531
Minke	1,200,537	26,446,186
Blue whale	4,371,463	27,158,938
Fin whale	2,668,509	27,549,771
Humpback	2,689,719	28,440,221
Gray whale A	1,591,221	24,091,848
Gray whale B	1,423,890	28,002,269

Table S4. Library and sequencing information for the hippo genome assembly. The estimate of the coverage is based on an assumed mammalian genome size of 3.1Gb.

Library	Insert size	No. Reads	No. reads used	Coverage
PE250	250 bp	771,298,964	619,136,098	19.8
PE500	500 bp	479,504,382	392,524,628	12.5
PE800	800 bp	384,369,550	305,871,802	9.77
MP2K	2 kb	343,921,854	28,532,3315	9.11
MP10K	10 kb	102,833,932	81,590,442	2.61

Table S5. Summary of repetitive elements in the hippopotamus genome. Repeats are grouped by their respective repeat family. Subfamily counts are given as subsets of their respective families. The table shows the number (count), combined lengths und percentage of the genome sequence. Here, repeat types are also given for subfamilies because the repeats of the hippotamus have not been characterized before.

Repeat type	# elements	lengths (in bp)	% sequence
SINE	877,948	140,668,128	5.80
Alu/B1	27	1,733	0.00
MIRs	530,250	72,759,825	3.00
LINEs	1,168,421	437,621,503	18.0
LINE1	744,824	332,612,025	13.7
LINE2	364,417	91,575,927	3.78
L3/CR1	46,665	10,136,404	0.42
RTE	11,571	3,127,613	0.13
LTR elements	431,887	152,603,906	6.29
ERVL	100777	43,336,314	1.79
ERVL-MaLRs	164,722	56,759,117	2.34
ERV class I	98,204	42,275,346	1.74
ERV class II	40,990	3,107,632	0.13
DNA elements	406,868	84,787,323	3.50
hAT-Charlie	221,833	42,833,726	1.77
TcMar-Tigger	72,285	19,831,649	0.82
Unclassified	7,031	1,301,848	0.05
Total interspersed repeats		816,982,708	33.7
Small RNA	33,7530	66,880,296	2.76
Satellites	125,788	52,266,573	2.15
Simple repeat	54,7104	22,624,259	0.93
Low complexity	102,688	5,155,551	0.21
Total masked			37.0

Table S6. A majority rule consensus analysis of 34,192 individual GF ML-trees. Only splits occurring more than 1% are shown. Species in order: 1. Grey whale (A) 2. Grey whale (B) 3. Blue whale 4. Sei whale (A) 5. Sei whale (B) 6. Humpback whale 7. Fin whale 8. Bowhead whale 9. North Atlantic right whale 10. Minke whale.

Set	Count	Frequency
Sets included in the consensus tree		
**.....	34192	1.000
*******..*	34192	1.000
...**.....	34190	1.000
.....**...	30755	0.899
..***.....	28662	0.838
*******...	19747	0.578
......	10315	0.302
Sets *not* included in the consensus tree		
..*****...	8918	0.261
*****.....	8721	0.255
..*****..*	3507	0.103
..***....*	3410	0.100
.....**..*	2737	0.080
**.......*	2204	0.064
.....*	2118	0.062
*****....*	1711	0.050
***.....	1408	0.041
......	1291	0.038
**...*....	991	0.029
...****...	985	0.029
.**...	915	0.027
..*..**...	907	0.027
***..**...	875	0.026
**....*...	848	0.025
..****....	595	0.017
..***.*...	562	0.016
...**....*	428	0.013
******....	416	0.012
*****.*...	410	0.012
..*......*	396	0.012

Note – The table summarizes the results from the consense analysis. The ranking is according to the number of occurrences of splits. Only splits occurring more frequent than 1% are shown. In each vertical column dots (.) and asterisks (*) represents one individual and its split into the respective group (. or *). For example: row one (**.....) has species 1 and 2 (both individuals of gray whale) as the most frequent split against all others, row two (*******..*) has species 8 and 9 (bowhead whale and NA right whale) splitting from the other with 34,192 occurrences.

Table S7. Common names, scientific names, accession numbers and source database of additional genomes that were included in the divergence time analyses.

Common name	Scientific Name	Accession	Source
Bajii	*Lipotes vexillifer*	GCF_000442215.1	RefSeq
Bottlenose dolphin	*Tursiops truncatus*	GCA_000151865.3	GenBank
Camel	*Camelus ferus*	GCF_000311805.1	RefSeq
Cow	*Bos taurus*	GCA_000003055.3	GenBank
Dog	*Canis lupus familiaris*	GCA_000002285.2	GenBank
Killer whale	*Orcinus orca*	GCF_000331955.2	RefSeq
Pig	*Sus scrofa*	GCA_000003025.4	GenBank
Sheep	*Ovis aries*	GCF_000298735.1	RefSeq
Sperm whale	*Physeter macrocephalus*	GCA_000472045.1	ENSEMBLE(pre)

Table S8. Calibration points used for the divergence time tree, node age estimates in million years agao (Ma) and references.

Node	Node age	Reference
Camelidae	63 - 73 Ma	Hasegawa et al. (2003)
Ruminantia	55 - 60 Ma	Arnason et al. (1996), Arnason and Gullberg (1996)
Hippopotamus	53.5 - 55 Ma	Bajpai and Gingerich (1998)
Cetacea	30.5 - 32.3 Ma	Mitchell (1989)
Mysticeti	<28 Ma	Fordyce (2002), Steeman et al. (2009)

Table S9. Divergence time estimates for Artiodactyla and Cetacea for nodes in the divergence time tree (Figure 5). The table shows the mean divergence times, 95% equal-tail confidence interval and 95% highest posterior density. For comparison divergence time estimates from Arnason et al. (2004) and McGowen et al. (2009) are given if present in the respective studies. Times are given in million years ago (Ma).

Node	Description	Mean	95% Equal-tail	95 % HPD	Arnason (2004)	McGowen (2009)
t_n25	Hippopotamidae	54.2	0.535 - 0.550	0.535 - 0.550	53.30	-
t_n26	Cetacea	31.8	0.307 - 0.324	0.308 - 0.324	35.00	36.36
t_n27	Odontoceti	27.6	0.212 - 0.310	0.226 - 0.315	32.1 $\pm$ 1.7	34.69
t_n28	Delphinidae + Lipotiidae	18.7	0.099 - 0.262	0.101 - 0.264	22.4 $\pm$ 1.5	24.70
t_n29	Delphinidae	8.81	0.035 - 0.165	0.030 - 0.158	-	10.80
t_n30	Mysticeti	28.3	0.275 - 0.295	0.274 - 0.294	20.90	28.79
t_n31	Balaenopteroidae	10.5	0.053 - 0.204	0.045 - 0.189	12.70	13.80
t_n32	Balaenopteridae ex. B. borealis	8.35	0.042 - 0.162	0.035 - 0.147	-	10.21
t_n33	Eschrichtiidae + Megaptera + B. physalus	7.49	0.036 - 0.148	0.031 - 0.135	-	9.04
t_n34	Megaptera + B. physalus	4.98	0.018 - 0.108	0.014 - 0.097	-	7.06
t_n35	Eschrichtius	0.08	0.000 - 0.002	0.000 - 0.002	-	-
t_n36	B. borealis + B. musculus	5.79	0.022 - 0.125	0.016 - 0.111	10.30	8.74
t_n37	B. borealis	0.29	0.001 - 0.008	0.001 - 0.007	-	-
t_n38	Balaenidae	4.38	0.012 - 0.126	0.007 - 0.103	-	5.38

References

Arnason, U. and Gullberg, A. (1996). Cytochrome b nucleotide sequences and the identification of five primary lineages of extant cetaceans. *Molecular biology and evolution*, 13(2):407–417.

Arnason, U., Gullberg, A., and Janke, A. (2004). Mitogenomic analyses provide new insights into cetacean origin and evolution. *Gene*, 333:27–34.

Arnason, U., Gullberg, A., Janke, A., and Xu, X. (1996). Pattern and timing of evolutionary divergences among hominoids based on analyses of complete mtDNAs. *Journal of Molecular Evolution*, 43(6):650–661.

Bajpai, S. and Gingerich, P. D. (1998). A new Eocene archaeocete (Mammalia, Cetacea) from India and the time of origin of whales. *Proceedings of the National Academy of Sciences*, 95(26):15464–15468.

Fordyce, R. E. (2002). Oligocene origins of skim-feeding right whales: A small archaic balaenid from New Zealand. *Journal of Vertebrate Paleontology*, 22(3):54A.

Hasegawa, M., Thorne, J. L., and Kishino, H. (2003). Time scale of eutherian evolution estimated without assuming a constant rate of molecular evolution. *Genes & Genetic Systems*, 78(4):267–283.

McGowen, M. R., Spaulding, M., and Gatesy, J. (2009). Divergence date estimation and a comprehensive molecular tree of extant cetaceans. *Molecular Phylogenetics and Evolution*, 53(3):891–906.

Mitchell, E. D. (1989). A New Cetacean from the Late Eocene La Meseta Formation Seymour Island, Antarctic Peninsula. *Canadian Journal of Fisheries and Aquatic Sciences*, 46(12):2219–2235.

Steeman, M. E., Hebsgaard, M. B., Fordyce, R. E., Ho, S. Y. W., Rabosky, D. L., Nielsen, R., Rahbek, C., Glenner, H., Sørensen, M. V., and Willerslev, E. (2009). Radiation of extant cetaceans driven by restructuring of the oceans. *Systematic Biology*, 58(6):573–585.

4.3 Retrophylogenomics in rorquals indicate large ancestral population sizes and a rapid radiation

Fritjof Lammers, Moritz Blumer, Cornelia Rücklé and Maria A. Nilsson[1]

Published in
Mobile DNA 10 (5) (2019).
10.1186/s13100-018-0143-2

Erklärung zu den Autorenanteilen an der Publikation
Was hat der Promovierende bzw. was haben die Koautoren beigetragen?

(1) zu Entwicklung und Planung
FL 80%, MAN 20 %

(2) zur Durchführung der einzelnen Untersuchungen und Experimente
Entwicklung Simulation: CR,FL (30 %); Durchführung Simulation: MB (10%);
NGS-Datenanalyse: FL (60 %)

(3) zur Erstellung der Datensammlung und Abbildungen
NGS-Datenaufbereitung, Mapping, Etablierung des Programms MELT, Auswertung der Simulation: MB (30 %); Abbildungen, Tabellen: FL (70%)

(4) zur Analyse und Interpretation der Daten
Analyse und Aufbereitung der MELT-Daten, Phylogenetische Analysen: FL (80 %);
Auswertung und Interpretation: FL, MB, MAN (20 %)

(5) zum Verfassen des Manuskripts
FL 70 %, MAN: 20 %, MB: 5%, CR: 5%

Zustimmende Bestätigungen der oben genannten Angaben:

_____ , _____	
Datum, Ort	Unterschrift Promovierender
_____ , _____	
Datum, Ort	Unterschrift Betreuer
_____ , _____	
Datum, Ort	ggfs. Unterschrift *corresponding author*

[1]Corresponding author

Lammers et al. Mobile DNA (2019) 10:5
https://doi.org/10.1186/s13100-018-0143-2

Mobile DNA

RESEARCH Open Access

Retrophylogenomics in rorquals indicate large ancestral population sizes and a rapid radiation

Fritjof Lammers[1,2,3], Moritz Blumer[1], Cornelia Rücklé[1] and Maria A. Nilsson[1,2*]

Abstract

Background: Baleen whales (Mysticeti) are the largest animals on earth and their evolutionary history has been studied in detail, but some relationships still remain contentious. In particular, reconstructing the phylogenetic position of the gray whales (Eschrichtiidae) has been complicated by evolutionary processes such as gene flow and incomplete lineage sorting (ILS). Here, whole-genome sequencing data of the extant baleen whale radiation allowed us to identify transposable element (TE) insertions in order to perform phylogenomic analyses and measure germline insertion rates of TEs. Baleen whales exhibit the slowest nucleotide substitution rate among mammals, hence we additionally examined the evolutionary insertion rates of TE insertions across the genomes.

Results: In eleven whole-genome sequences representing the extant radiation of baleen whales, we identified 91,859 CHR-SINE insertions that were used to reconstruct the phylogeny with different approaches as well as perform evolutionary network analyses and a quantification of conflicting phylogenetic signals. Our results indicate that the radiation of rorquals and gray whales might not be bifurcating. The morphologically derived gray whales are placed inside the rorqual group, as the sister-species to humpback and fin whales. Detailed investigation of TE insertion rates confirm that a mutational slow down in the whale lineage is present but less pronounced for TEs than for nucleotide substitutions.

Conclusions: Whole genome sequencing based detection of TE insertions showed that the speciation processes in baleen whales represent a rapid radiation. Large genome-scale TE data sets in addition allow to understand retrotransposition rates in non-model organisms and show the potential for TE calling methods to study the evolutionary history of species.

Keywords: Evolution, Phylogenetics, Whales, Transposable elements, Retrotransposon

Background

The bifurcating tree of life, where at each speciation event one ancestral lineage split into two new species, is a concept deeply rooted in the field of evolutionary biology. The opposite, that several new lineages diverge from the same speciation event, a so called polytomy, is mostly regarded as an artefact of limited phylogenetic information [1]. The sequencing and analyses of complete genomes was expected to finally resolve ambiguous relationships by providing enormous amounts of data [2]. Instead of resolving long standing phylogenetic controversies, genome-scale datasets revealed a lot of natural complexity in the phylogenetic data that previously had been deemed as noise [3, 4].

The evolutionary history of baleen whales (Mysticeti) is a prominent example of a phylogeny that lacked a scientific consensus for a long time [5–8]. In particular, the relationships among rorquals (Balaenopteridae) and gray whales (Eschrichtiidae) were contentious. While some studies showed that the only extant species of gray whales (*Eschrichtius robustus*) is phylogenetically placed within rorquals [6–8], others placed the gray whale as a sister group to rorquals, which was expected given its different morphology and feeding behaviour [5, 9].

* Correspondence: maria.nilsson-janke@senckenberg.de
[1]Senckenberg Biodiversity and Climate Research Centre, Senckenberg Gesellschaft für Naturforschung, Senckenberganlage 25, 60325 Frankfurt am Main, Germany
[2]LOEWE Centre for Translational Biodiversity Genomics (LOEWE-TBG), Senckenberganlage 25, 60325 Frankfurt am Main, Germany
Full list of author information is available at the end of the article

© The Author(s). 2019 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Recently, whole-genome sequencing (WGS) of nearly all extant baleen whale species suggested that the rapid radiation of rorquals might represent a hard polytomy [10]. To further explore if the baleen whale phylogeny contains a polytomy, we use transposable element (TE) insertions. TEs are a robust and independent type of phylogenetic markers, that overcomes many limitations of sequence based phylogenetics, i.e. based on single nucleotide variants (SNV) [11]. Furthermore, TEs evolve neutrally and occur interspersed throughout the genome. Hence, they avoid potentially biased phylogenetic signals from gene tree error or linkage disequilibrium that can occur in sequence-based multi-locus analyses [12]. In addition, TE insertions are virtually homoplasy-free because parallel insertions in the large genomic space are very rare [11]. Also, they are less prone to reversals or mutational saturation that can affect SNV-based phylogenetic inference [11].

In baleen whale genomes, the most abundant TEs are short and long interspersed nuclear elements (SINEs and LINEs), covering 24.5% of the bowhead whale genome [10, 13]. The most abundant SINE family in baleen whales are CHR2 elements, which are named after their presence in **C**etacea, **H**ippopotamidae and **R**uminants [14] and emerged at least 56 million years ago (Mya). Like most other SINEs, the non-autonomous CHR2 elements are derived from a tRNA sequence. They are mobilized by the enzymatic machinery of LINE1 elements via an RNA intermediate that is reverse transcribed to cDNA and reintegrated into the genome. Compared to LINEs, their relatively high insertion frequencies make SINEs ideally suited for phylogenetic inference in mammalian genomes [11]. TEs have a long history of being used as phylogenetic markers for different cetacean groups [15–17].

Due to advances in genome sequencing and software development thousands of TE insertions can be inferred from multiple genomes across species and individuals [18, 19]. Thus, genome-scale TE detection was successfully applied to analyze retrotransposition in several vertebrate clades outside humans [20–23]. Furthermore, WGS based approaches proved extremely valuable in phylogenetic inference because they can increase the number of discovered TE insertions a thousand-fold, providing enhanced statistical power and the possibility to detect processes of reticulate evolution [23]. By contrast, PCR-based approaches have relied on tedious and time-consuming experimental work to find a few dozens of phylogenetically informative TE insertions from hundreds to thousands of candidate loci [24, 25]. Selection of candidate loci using an experimental approach was often based on a single genome sequence, introducing an ascertainment bias in the phylogenetic signal [17, 26, 27] that can be avoided by the use of large scale WGS sequencing and bioinformatic pipelines.

Here, we identified 91,859 CHR2 insertions in the available baleen whale genomes. This dataset was used to reconstruct the rorqual species tree and allowed us to quantify evolutionary conflict originating from their rapid radiation that took place approximately 8 Mya, coinciding with the onset of modern global oceanic circulation.

Results

WGS mapping and TE variation discovery

We mapped 11 WGS datasets from baleen whales with a coverage depth between 7 and 30 X to the bowhead whale (*Balaena mysticetus*) genome sequence [13] (Additional file 1: Table S1). From the mapped data, the Mobile Element Locator Tool (MELT) [19] called 488,373 non-reference (i.e. absent from the bowhead whale genome) CHR2 insertions, of which 327,488 (67.1%) passed stringent quality filtering. The bowhead whale is a natural outgroup to rorquals and gray whales, hence we focused on calling non-reference insertions in the 11 baleen whales to obtain an ascertainment bias free marker set for rorquals and gray whales. The total number of extracted CHR2 insertion calls per species ranged between 27,994 and 38,182, except for the North Atlantic right whale (*Eubaleana glacialis*), for which 6608 were found (Table 1). The North Atlantic right whale diverged from the bowhead whale about 4.4 Mya, hence fewer variable CHR2 loci reflect a closer genetic distance. In comparison, the divergence time of right whales and the bowhead whale to rorquals and gray whales is ~ 28 Ma. For clarity, we follow the nomenclature by ref. 10 to include the gray whale within rorquals sensu lato (Balaenopteridae + Eschrichtiidae).

Table 1 Numbers of all CHR2 insertion calls, as well as the amount of heterozygous insertions (Het) in baleen whale genomes compared to the bowhead whale genome

Sample	No CHR2 calls	Het
Blue whale	37,133	26,942
Fin whale	27,994	13,712
Gray whale (eastern) A	36,064	14,648
Gray whale (eastern) B	38,182	17,449
Gray whale (western) A	32,057	24,922
Gray whale (western) B	32,735	22,544
Humpback whale	28,618	14,622
Minke whale	28,606	12,089
North Atlantic right whale	6608	4221
Sei whale A	29,874	11,242
Sei whale B	29,617	11,079
Total	327,488	173,470

Extensive simulations to test the performance of MELT on our dataset showed that a sequencing depth of 5 X or higher is sufficient to reach true positive rates (TPR) of 99% for CHR2 insertions (Additional file 1: Figure S1A). Similarly, 92% of called CHR2 insertions were correctly recognized as homozygous indicating a high genotype accuracy on our dataset (Additional file 1: Figure S1B). MELTs internal filtering reduced sensitivity slightly (Additional file 1: Figure S1C, D), however, our simulations showed that the most effective filters affected all mapped genomes equally because they were based on properties of the reference genome, e.g. the presence of low-complexity regions (Additional file 1: Figure S2). Hence, these filters are not expected to create biases between samples that would influence phylogenetic inference. Furthermore, MELT-Split, which jointly genotypes all genomes, highly improved the detection of orthologous insertions compared to analyzing each genome individually and later combining the results. In summary, the simulations showed that our approach generated a dataset of high-quality baleen whale TE insertions with the corresponding orthology information that are suitable for evolutionary analyses.

TE phylogenomics recovers rorqual speciation history

By creating a presence-absence matrix from 327,488 genotyped CHR2 insertion sites in all genomes, 91,859 orthologous integration events were identified that took place during the evolution of baleen whales. Based on the presence-absence matrix, phylogenetic trees were reconstructed using Dollo parsimony, Bayesian inference (BI), and Neighbor-Joining (NJ) methods. The three reconstruction methods indicated a common monophyletic origin of Balaenopteridae and Eschrichtiidae (Fig. 1a, Additional file 1: Figure S3) and placed the gray whale as the sister species to the fin whale (*Balaenoptera physalus*) and humpback whale (*Megaptera novaeangliae*) clade. The minke whale (*Balaenoptera acutorostrata*) was reconstructed as the most basal rorqual species. In the NJ and BI trees, blue whale (*Balaenoptera musculus*) and sei

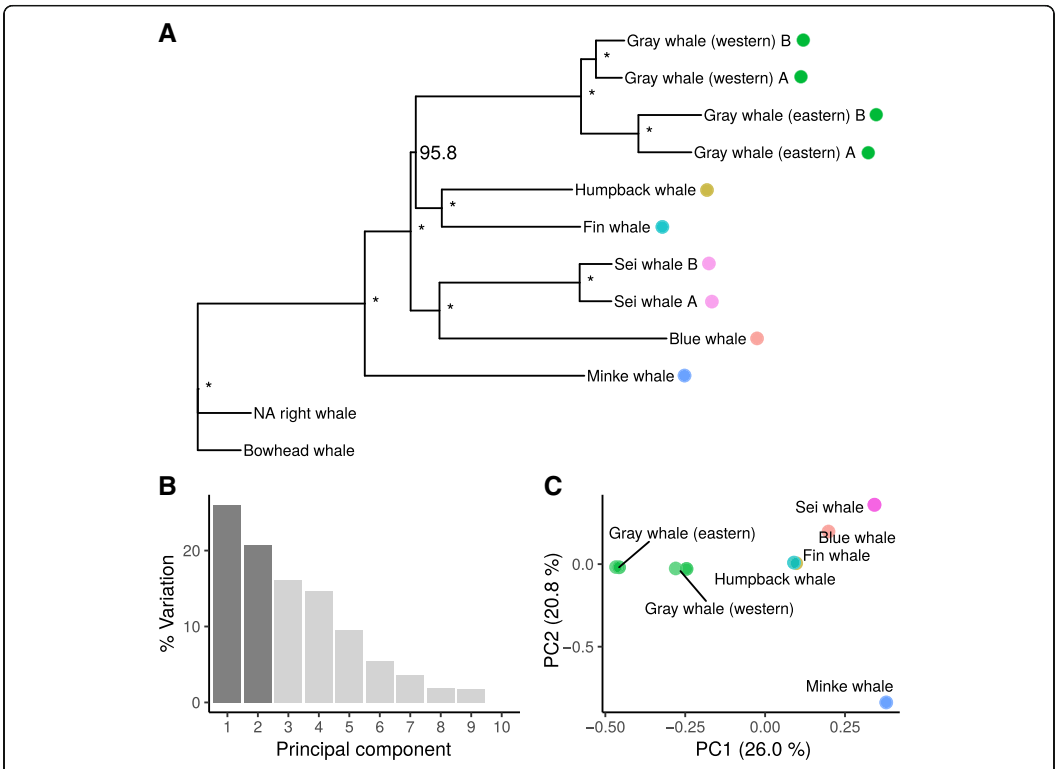


Fig. 1 Phylogenetic signal calculated from 91,859 CHR2 insertions in baleen whales. **a** Neighbor-Joining tree based on CHR2 insertions. All nodes received bootstrap values of 95% or higher (100% shown as asterisk). **b** Percentage of variation explained by principal components 1–10 in the PCA. **c** Scatterplot of the first two principal components (PC1 and PC2) among baleen whale genomes

whales (*Balaenoptera borealis*) formed a monophyletic clade as a sister group to the fin, humpback and gray whales. The CHR2 Dollo parsimony tree differed slightly from this topology because it reconstructed blue and sei whale as two separate lineages outside the fin, humpback and gray whale clade (Additional file 1: Figure S3 A). All trees received high node support with bootstrap values > 0.95 (Dollo parsimony, NJ) and 100% posterior probabilities (BI).

Although these tree reconstruction methods can by design only yield bifurcating topologies and cannot take conflicting genomic signals into account, considerable amount of phylogenetic conflict is indicated by low consistency indices (CI) (ranging between 0.629 and 0.646). The CI is a measure for tree support that indicates the fraction of minimum character changes compared to the observed number of changes, i.e. the tree length. If all character changes are consistent with the reconstructed tree, the CI is 1.0.

Analyzing the phylogenetic signal from CHR2 insertions among rorquals sensu lato using a principal component analysis (PCA) resulted in only the minke whale being clearly separated from the other species in the first two components, which together explained more than 50% of the variance in the dataset (Fig. 1b and c). While most species were found to be distinct along the first component, gray, fin and humpback whale were nearly indistinguishable on the second component. Furthermore, on the second component, the intraspecific differentiation between the two gray whale populations was as high as between other species pairs (Fig. 1c).

Network analysis reveals phylogenetic conflict

The low CIs of the phylogenetic trees indicate considerable amounts of phylogenetic conflict in the baleen whale genomes. To further explore these evolutionary signals, a median-joining network was calculated in order to uncover signals that otherwise remain hidden by traditional bifurcating tree-reconstruction algorithms. The phylogenetic network of CHR2 insertions showed a star-like web in the center of Balaenoptera and Eschrichtiidae (rorquals sensu lato) (Fig. 2a). Edges in the network that cluster the

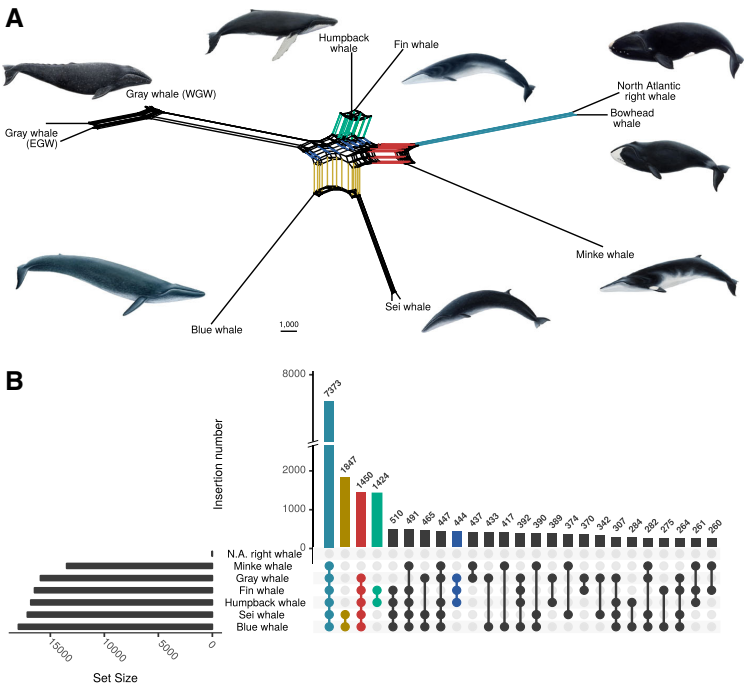


Fig. 2 Phylogenetic conflict among baleen whales inferred from CHR2 insertions. **a** Phylogenetic median-joining network based on 91,859 CHR2 insertions. **b** Distribution of phylogenetic signals in the dataset. Each synapomorphic CHR2 insertion is considered a phylogenetic signal for the common ancestry for the taxa carrying the insertions. The x-axis shows synapomorphic CHR2 insertions between species listed on the left-hand side. Bars on the y-axis show the number of insertions for the respective synapomorphies. The set sizes on the left-hand side show the total number of insertions present per species. Whale paintings are by Jon Baldur Hildberg (www.fauna.is)

gray whale with either the blue and sei whales and/or fin and humpback whales had similar lengths, thus indicating equally strong phylogenetic signal for both topologies.

A quantification of shared CHR2 insertions in baleen whales showed that the four strongest phylogenetic signals support the NJ tree (Fig. 2b) and are in agreement with the evolutionary history of rorquals inferred from genomic sequence analyses [10]. For example, the strongest signal consisted of 7373 synapomorphic CHR2 insertions shared by all rorquals sensu lato and supports a common ancestry of this clade. Within rorquals, 1450 insertions support that the gray whale diverged after the minke whale, confirming the paraphyly of rorquals sensu stricto. The monophyly of blue and sei whale as well as of fin and humpback whale was supported by 1847 and 1424 insertions, respectively. These strong signals match the well supported nodes in the reconstructed phylogenetic tree (Fig. 1a): the minke whale is clearly distinct from the other rorquals, and the sister group relationships of blue and sei whale as well as of fin and humpback whale are strongly supported. In contrast to other phylogenetic signals incongruent to the species tree, the numbers of TE insertions for the different phylogenetic positions of the gray whale among rorquals are highly similar and make a differentiation between evolutionary scenarios difficult. A ratio of 510:465:444 CHR2 insertions place the gray whale outside a fin, humpback, blue and sei whale clade (510), as sister clade to blue and sei whale (465) or as sister clade to fin and humpback whale (444), respectively (Additional file 1: Figure S4). Hence, this speciation event in the phylogenetic tree appears intuitively as unresolved and in fact a polytomy was only marginally rejected by the KKSC bifurcation test (p = 0.0204) [26]. In addition, a plethora of alternative phylogenetic signals of similar strengths illustrate the star-like radiation of Balaenopteridae and Eschrichtiidae. For example, the gray whale shares 433, 374 and 370 CHR2 insertions exclusively with the blue, humpback and fin whale, respectively. With regard to the previously established species tree, these insertions appear to be signals for ILS, however, they can not be considered by the KKSC test [26]. The KKSC test updates the statistical framework introduced by Waddell et al. [28] to test for the significance of conflicting phylogenetic signals from TE insertions to distinguish between ILS and introgression scenarios.

TE insertion dynamics

To explore the insertion dynamics of CHR2 in baleen whales, we investigated the genetic diversity and the insertion rates across time. We mapped the insertion points of all 91,859 CHR2 insertions on the baleen whale species tree [10] and calculated the frequency of heterozygous insertions on basis of the genotyping information provided by MELT. This allowed us to track how many insertions from each ancestral branch were fixed over time. Not surprisingly, several terminal branches exhibit high rates of heterozygous CHR2 insertions such as the two gray and sei whale populations and the blue whale (Additional file 1: Figure S5). High rates of heterozygous insertions originate also from the ancestral branches that led to the ancestor of gray, fin, humpback, sei and blue whales as well as from the ancestral branch to the fin, humpback and gray whale clade. The genomic heterozygosity of CHR2 insertions was lower in the sei whale branch and the fin and humpback whale clades, branches that exhibit less phylogenetic conflict (Fig. 2).

CHR2 insertion rates were calculated by mapping the insertion numbers on the species tree and using previously estimated divergence times [10] and an average generation time of 24.4 years for extant baleen whales [29]. The estimated insertion rates were relatively stable across the evolutionary lineages and ranged between 0.013–0.138 CHR2 insertions per generation (Additional file 1: Figure S6). The insertion rates at the terminal and shallow branches were relatively low and varied between 0.013 and 0.035. For the ancestral branch to gray, fin, humpback, blue and sei whale a ~ 10-fold increase in insertion rate was observed compared to other branches. The majority of CHR2 insertions that occured on this branch are incongruent to the bifurcating species tree. Repeat landscapes of minke and bowhead whale genome assemblies illustrate the evolution of TE sequences over time, by plotting the frequencies of sequence divergence to the TE consensus sequences. Both whale species show an increase in frequency of low-divergent SINEs (5–10% CpG-adjusted divergence), that could indicate an amplification burst of these elements (Additional file 1: Figure S7). The presence of a similar peak in both species at the same divergence indicate it must have occurred before their divergence at ~ 28 Mya.

Discussion

Here we have performed the first genome-scale analysis of TE insertions in whales based on next-generation sequencing technology. The included dataset, consisting of 91,859 insertion events across eight baleen whale species, exceeds the dataset size from a previous experimental approach by several magnitudes [16]. Our dataset made it possible to reconstruct the baleen whale evolutionary history and a detailed quantification of phylogenetic conflict.

Many previous studies have attempted to resolve the phylogeny of baleen whales and to clarify the evolutionary origin of the gray whale (family Eschrichtiidae). The gray whale is ecomorphologically derived from the family Balaenopteridae [5, 9] because it is the only bottom-feeding species within a clade of strictly lunge-feeding species [30] leading to confusion about its taxonomic position among baleen whales. Using TEs as virtually homoplasy-free and independent phylogenetic markers overcomes limitations from single-nucleotide based phylogenies [11] and should

provide a more detailed understanding about the evolution of baleen whales. Thus, we expected that a detailed analysis of TE insertions would finally settle the baleen whale relationships and also add additional information about the rate of retrotransposition in the slowest evolving mammals.

An evolutionary network analysis together with a detailed analysis of phylogenetically incongruent CHR2 insertions suggests that the speciation of rorquals represents a divergence that might not be entirely dichotomous. This is in spite that the TE based phylogenies were well supported and highly identical to the multi-locus coalescent tree generated from 34,192 sequence based gene trees [10] and a supermatrix tree [7]. However, careful interpretation is warranted given that bootstrap support and posterior probability were designed to assess sampling error of single genes, not genome-scale datasets and might lead to wrong conclusions about the species relationships [31]. Using bootstrap replicates and Bayesian probabilities to infer branch support is common practice, however, well-supported branches might merely be the result of an oversimplified evolutionary model if the dataset is large and the phylogenetic signal is not tree-like. Our in-depth analysis of conflicting synapomorphic TE insertions in baleen whale genomes show that the high statistical support in the phylogenetic trees is based on marginal numeric differences. Unfortunately, methods and models to reconstruct phylogenies from genome-scale multi-locus TE insertion datasets are not as developed as for nucleotide substitutions.

The presence of several equally strong conflicting phylogenetic signals in the CHR2 dataset can be caused by a) insufficient character sampling leading to an unresolved divergence (soft polytomy), b) near-instantaneous speciation and subsequent incomplete lineage sorting (ILS), or c) speciation under genetic exchange. Given the data presented here, it is highly unlikely that the divergence of the gray whale and its sister lineages represent a soft polytomy (a), as our extensive dataset of 91,859 CHR2 insertions is distributed across the near complete 2.3 Gb genome sequence of baleen whales and each node in the phylogeny is supported by several hundred insertions (Fig. 2b). In addition, a confounding effect from incorrect phylogenetic signal is marginal because SINE insertions are virtually free from homoplasy.

ILS (b) is the persistence of ancient polymorphisms across speciation events and has been observed in several TE-based phylogenomic studies [32–34], including a study investigating baleen whale relationships [16]. Several factors, such as a rapid radiation, large or expanding ancestral effective population sizes (N_e) and consequently a slow evolutionary fixation rate favor the occurrence of ILS [33]. The gray whale and the ancestors of the blue- plus sei whales and fin- plus humpback whales rapidly diverged from each other within less than one million years, as is evident from the star-like phylogenetic network (Fig. 2a) and previous divergence time estimates [10, 35]. In addition, a large ancestral N_e is suggested by the high number of species-tree incongruent CHR2 insertions and the large fraction of evolutionary old and still unfixed, heterozygous insertions that integrated on the ancestral branches with the highest degree of ILS (Additional file 1: Figure S5, and S6). The genome-wide analysis of CHR2 insertion thus strongly indicates that the ancestral rorqual population exhibited large population sizes and radiated rapidly. Also, explicit modeling of the demographic histories of baleen whales based on genomic data indicates large ancestral population sizes of whales [10]. However, these estimates do not reach back enough in time to cover the timeframe of the radiation.

Whales are the largest living animals and known for their slow physiological and evolutionary rate [36]. They exhibit the slowest nucleotide substitution rate among mammals, estimated to be 10 times slower than among primates [37]. Our estimates indicate that the rate of SINE insertions is about 50% slower than in humans, for which a mean rate of 0.046 Alu insertions per generation per genome was estimated [38]. However, we also observe a 10-fold increased CHR2 insertion rate on the branch to the fin, humpback, gray, blue and sei whale clade. Similar strong fluctuations in SINE insertion rates across evolutionary time, like estimated within baleen whales, were also reported for great apes [20].

Finally, a potential third cause for a conflicting phylogenetic signal (c) is that the emerging whale species might have exchanged genetic material for a long time because vicariance is more difficult to maintain in the marine than in the terrestrial environment. Hence, also speciation with genetic exchange of baleen whales might have caused trans-species polymorphisms [10, 39]. Whether the resulting genomic mosaicism is a result of speciation with genetic exchange or from ILS is however not possible to determine [40] and both processes are plausible for baleen whales. Either process or a combination of both could have created the observed phylogenetic signals that are incompatible with a strictly bifurcating tree. More detailed investigation of these processes require new methods that examine patterns of phylogenetic signals from TE insertions with respect to speciation processes and gene flow.

Conclusions

This study demonstrates the suitability of WGS datasets to infer TE insertions, one of the largest contributor to genomic variation in mammals [41]. Thus, TE insertions are a highly valuable source for

comparative genomics and for reconstructing phylogenies. In line with the first application of TE-based phylogeny of baleen whales [16] and a recent nucleotide-based study [10], the radiation of rorquals sensu lato appears to represent a hard polytomy when depicted as a phylogenetic tree because alternative phylogenetic scenarios are equally well supported. Therefore, a better representation of the rorquals' evolutionary history would be to represent the divergences in a phylogenetic network [10], allowing for the incorporation of ILS and genetic exchange between species as horizontal reticulations. We anticipate that a population-wide sampling of baleen whales might illuminate the divergence processes in more detail.

Materials and methods

WGS mapping

Whole-genome sequencing data from ref. 10 plus additional samples of two gray whales and a fin whale [42, 43] were quality-checked with FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), trimmed if necessary with Trimmomatic [44] and mapped to the bowhead whale genome with BWA [45] (Additional file 1: Table S1). The bowhead whale (*Balaena mysticetus*) genome assembly [13] was chosen for reference mapping over the more continuous minke whale genome because it is a natural outgroup to the rorqual species and thus eliminates TE detection bias between samples [23].

TE detection

The Mobile Element Locator Tool (MELT) [19] was run in the Split mode on all scaffolds larger than 100 kb. A consensus file for TE detection was created according to the MELT manual. We chose the general consensus sequence of the CHR2 SINE family, that was active during the evolution of Cetacea [46]. Seven different subfamilies of CHR2 have been described for cetaceans [47], that contain indels compared to the general CHR2 consensus sequence. Using the full length general consensus of CHR2 [14] and allowing for 10% mismatches makes a broader detection of CHR2 insertions in MELT possible. To annotate all copies of the CHR SINE family elements in the bowhead whale genome, the genome sequence was repeat-masked (http://www.repeatmasker.org/) with the Cetartiodactyla repeat library. BEDOPS [48] converted the RepeatMasker output into BED format.

Simulation and sensitivity analysis

Prior to TE calling, we performed a sensitivity and specificity analysis using our custom-made TE calling assessment pipeline ESAT (Element Simulation Analysis Tool) using sequences and parameters matching our whale dataset. We selected the longest scaffold (5 Mb) from the bowhead whale assembly to serve as a sample genome for our sensitivity analysis. We randomly integrated 200 CHR2 SINEs in the sample genome sequence and simulated paired-end Illumina reads from the resulting sequence with SimSeq (https://github.com/jstjohn/SimSeq) at sequencing coverage levels ranging from 1 to 30 X coverage. For read simulation we generated an error-profile typical for our whale resequencing datasets. Reads were mapped to the sample genome with BWA [45] as described above and MELT was used to call the CHR2 SINE insertions from our simulated genome. We generated 10 replicates per simulation. To analyze the performance of MELT, we assessed if the detected non-reference TE insertions matched the simulated TE locations using BEDtools [49]. The detection rate (DETR) reflects the sensitivity of MELT to successfully identify a TE insertion. True positive rate (TPR), false positive rate (FPR) and false negative rates (FNR) were calculated from the detected TEs to estimate MELT's accuracy on the whale dataset. Finally, the proportion of correctly genotyped insertions among the detected variants was recorded. We made ESAT publicly available on https://github.com/crueckle/ESAT.

Phylogenomic analysis

Orthologous TE insertion calls across the taxon sampling were identified using the GroupAnalysis and Genotype algorithms in MELT. TE insertion calls passing internal MELT filters were extracted with bcftools filter (www.htslib.org). A NEXUS-formatted presence absence matrix of orthologous TE insertions was created with a modified version of vcf2phylip [50]. Phylogenies were reconstructed using Neighbor-Joining and Dollo Parsimony in PAUP* [51]. Under Dollo Parsimony, only character state changes from absence to presence (0 to 1) are allowed, thus matching the evolutionary model of TE insertions. Heuristic tree search was conducted with random addition of sequences and 100 repetitions using Tree Bisection and Reconnection (TBR) as branch swap algorithms. Bootstrap support values were calculated from 1000 replicates. Likelihood scores for each tree were calculated using the 'lscores' command. A Bayesian inference tree was calculated in MrBayes v.3.2.6 [52] using "irreversible" character type (ctype irreversible:all) with 10e7 generations and sampling every 1000th generations, 25% of the samples were discarded as burn-in. Principal component analysis (PCA) for the filtered CHR2 datasets were conducted with the SNPRelate package for R. Phylogenetic median joining networks were generated in SplitsTree4 [53]. The intersection diagram was created with UpSetR [54]. For gray and sei whales, only TE insertions present in all individuals of the respective species were considered.

Insertion rates

Per-branch insertion rates were calculated from the number of CHR2 insertions that we had mapped to the species tree from ref. 10. This tree was used because it is the best available bifurcating representation of the baleen whales evolutionary history and is congruent with other recent studies on baleen whale phylogeny [7]. Species-tree incongruent CHR2 insertions were assumed to be the result of ILS and accordingly mapped to the most recent ancestral branch leading to the affected species. The insertion rate was calculated by the equation $\mu = \eta_{CHR2}{}^{*}b/24.4$ with n_{CHR2} for the number of CHR2 insertions and b as the branch length in years. The mean generation time of 24.4 years was calculated for from recent generation time estimates of the studied species [29].

Additional files

Additional file 1: Table S1. List of samples with accession numbers and sequencing properties. **Figure S1.** Simulation results for CHR2 detection with MELT at varying depth of coverage using dataset specific parameters. **Figure S2** Frequency of filters applied by MELT to exclude low-quality CHR2 calls. **Figure S3** Phylogenetic trees of baleen whales reconstructed with CHR2 insertions. A) Dollo-Parsimony tree reconstructed in PAUP*. Asteriks indicate 100 % bootstrap support (500 replicates), lower bootstrap support is given as numbers. B) Bayesian inference tree with posterior probability given for nodes. **Figure S4** Three alternative relationships in the rorqual radiation and the number of CHR2 insertion that support them. **Figure S5** Phylogenetic tree of rorquals with frequency of heterozygous insertions per branch. **Figure S6** CHR2 insertion rates per generation. **Figure S7** Repeat landscapes of minke whale and bowhead whale based on available assemblies. (PDF 475 kb)

Additional file 2: Data S1: VCF file with filtered CHR2 variants in baleen whales called by MELT. (ZIP 10547 kb)

Additional file 3: Data S2: NEXUS file with the presence-absence matrix of CHR2 insertions in baleen whales encoded as 1 (presence) and 0 (absence). (ZIP 159 kb)

Abbreviations

BI: Bayesian inference; CI: Consistency index; DETR: Detection rate; ESAT: Element simulation analysis tool; FNR: False negative rate; FPR: False positive rate; ILS: Incomplete lineage sorting; MELT: Mobile element locator tool; Mya: Million years ago; N_e: Effective population size; NJ: Neighbor-Joining; PCA: Principal component analysis; SNV: Single nucleotide variant; TBR: Tree bisection and reconnection; TE: Transposable element; TPR: True positive rate; WGS: Whole genome sequencing

Acknowledgements

We acknowledge the constructive help from Eugene Gardner (http://melt.igs.umaryland.edu/) in running MELT on our dataset. We thank Axel Janke and the members of the working group for valuable comments to the manuscript and fruitful discussions. Whale paintings in Fig. 2 are by Jon Baldur Hildberg (www.fauna.is).

Funding

This work was funded by the Senckenberg Gesellschaft für Naturforschung, a member of the Leibniz Association. The present manuscript is a result of the Centre for Translational Biodiversity Genomics (LOEWE-TBG) and was supported through the programme "LOEWE – Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz" of Hesse's Ministry of Higher Education, Research, and the Arts.

Availability of data and materials

All sequencing data is accessible from the respective accession numbers listed in Additional file 1: Table S1. The presence absence data (Additional file 3: Data S2) and the insertion calls (Additional file 2: Data S1) are provided in Additional file 1: Table S1, respectively. The modified version of vcf2phylip is available at https://github.com/mobilegenome/vcf2phylip.

Authors' contributions

FL and MN conceived the study. MB and FL performed the analyses. CR wrote the simulation pipeline (ESAT). FL, MB and MN interpreted the results. FL wrote the manuscript with input from MN. All authors read, gave comments and helped to revise the final version of the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Senckenberg Biodiversity and Climate Research Centre, Senckenberg Gesellschaft für Naturforschung, Senckenberganlage 25, 60325 Frankfurt am Main, Germany. [2]LOEWE Centre for Translational Biodiversity Genomics (LOEWE-TBG), Senckenberganlage 25, 60325 Frankfurt am Main, Germany. [3]Institute for Ecology, Evolution and Diversity, Goethe University Frankfurt, Biologicum, Max-von-Laue-Straße 13, 60439 Frankfurt am Main, Germany.

Received: 23 October 2018 Accepted: 18 December 2018
Published online: 21 January 2019

References

1. Hoelzer GA, Meinick DJ. Patterns of speciation and limits to phylogenetic resolution. Trends Ecol Evol. 1994;9:104–7.
2. Murphy WJ, Pevzner PA, O'Brien SJ. Mammalian phylogenomics comes of age. Trends Genet. 2004;20:631–9.
3. Delsuc F, Brinkmann H, Philippe H. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet. 2005;6:361–75.
4. Bapteste E, van Iersel L, Janke A, Kelchner S, Kelk S, McInerney JO, et al. Networks: expanding evolutionary thinking. Trends Genet. 2013;29:439–41.
5. Árnason U, Gullberg A, Janke A. Mitogenomic analyses provide new insights into cetacean origin and evolution. Gene. 2004;333:27–34.
6. Hassanin A, Delsuc F, Ropiquet A, Hammer C, Jansen Van Vuuren B, Matthee C, et al. Pattern and timing of diversification of Cetartiodactyla (Mammalia, Laurasiatheria), as revealed by a comprehensive analysis of mitochondrial genomes. Comptes Rendus - Biologies. 2012;335:32–50.
7. Marx FG, Fordyce RE. Baleen boom and bust: a synthesis of mysticete phylogeny, diversity and disparity. R Soc Open Sci. 2015;2:140434.
8. Rychel AL, Reeder TW, Berta A. Phylogeny of mysticete whales based on mitochondrial and nuclear data. Mol Phylogenet Evol. 2004;32:892–901.
9. Gatesy J, Geisler JH, Chang J, Buell C, Berta A, Meredith RW, et al. A phylogenetic blueprint for a modern whale. Mol Phylogenet Evol. 2013;66:479–506.
10. Árnason Ú, Lammers F, Kumar V, Nilsson MA, Janke A. Whole-genome sequencing of the blue whale and other rorquals finds signatures for introgressive gene flow. Sci Adv. 2018;4:eaap9873.
11. Shedlock AM, Okada N. SINE insertions: powerful tools for molecular systematics. BioEssays. 2000;22:148–60.
12. Edwards SV, Xi Z, Janke A, Faircloth BC, McCormack JE, Glenn TC, et al. Implementing and testing the multispecies coalescent model: a valuable paradigm for phylogenomics. Mol Phylogenet Evol. 2016;94:447–62.
13. Keane M, Semeiks J, Webb AE, Li YI, Quesada V, Craig T, et al. Insights into the evolution of longevity from the bowhead whale genome. Cell Rep. 2015;10:112–22.

14. Shimamura M, Yasue H, Ohshima K, Abe H, Kato H, Kishiro T, et al. Molecular evidence from retroposons that whales form a clade within even-toed ungulates. Nature. 1997;388:666–70.
15. Nikaido M, Matsuno F, Abe H, Shimamura M, Hamilton H, Matsubayashi H, et al. Evolution of CHR-2 SINES in cetartiodactyl genomes: possible evidence for the monophyletic origin of toothed whales. Mamm Genome. 2001;12:909–15.
16. Nikaido M, Hamilton H, Makino H, Sasaki T, Takahashi K, Goto M, et al. Baleen whale phylogeny and a past extensive radiation event revealed by SINE insertion analysis. Mol Biol Evol. 2006;23:866–73.
17. Nikaido M, Piskurek O, Okada N. Toothed whale monophyly reassessed by SINE insertion analysis: the absence of lineage sorting effects suggests a small population of a common ancestral species. Mol Phylogenet Evol. 2007;43:216–24.
18. Ewing AD. Transposable element detection from whole genome sequence data. Mob DNA. 2015;6:24.
19. Gardner EJ, Lam VK, Harris DN, Chuang NT, Scott EC, Pittard WS, et al. The Mobile element locator tool (MELT): population-scale mobile element discovery and biology. Genome Res. 2017;27:1916–29.
20. Hormozdiari F, Konkel MK, Prado-Martinez J, Chiatante G, Herraez IH, Walker J a, et al. Rates and patterns of great ape retrotransposition. Proc Natl Acad Sci. 2013;110:13457–62.
21. Ruggiero RP, Bourgeois Y, Boissinot S. LINE insertion polymorphisms are abundant but at low frequencies across populations of *Anolis carolinensis*. Front Genet. 2017;8:1–14.
22. Suh A, Smeds L, Ellegren H. Abundant recent activity of retrovirus-like retrotransposons within and among flycatcher species implies a rich source of structural variation in songbird genomes. Mol Ecol. 2018;27:99–111.
23. Lammers F, Gallus S, Janke A, Nilsson MA. Phylogenetic conflict in bears identified by automated discovery of transposable element insertions in low-coverage genomes. Genome Biol Evol. 2017;9:2862–78.
24. Churakov G, Kriegs JO, Baertsch R, Zemann A, Brosius J, Schmitz J. Mosaic retroposon insertion patterns in placental mammals. Genome Res. 2009;19:868–75.
25. Nilsson M a, Churakov G, Sommer M, Tran NV, Zemann A, Brosius J, et al. Tracking marsupial evolution using archaic genomic retroposon insertions. PLoS Biol. 2010;8:e1000436.
26. Kuritzin A, Kischka T, Schmitz J, Churakov G. Incomplete lineage sorting and hybridization statistics for large-scale retroposon insertion data. PLoS Comput Biol. 2016;12:e1004812.
27. Dodt WG, Gallus S, Phillips MJ, Nilsson MA. Resolving kangaroo phylogeny and overcoming retrotransposon ascertainment bias. Sci Rep. 2017;7:16811.
28. Waddell PJ, Kishino H, Ota R. A phylogenetic foundation for comparative mammalian genomics. Genome Inform Ser. 2001;154:141–54.
29. Taylor BL, Chivers SJ, Larese J, Perrin WF. Generation length and percent mature estimates for IUCN assessments of cetaceans. La Jolla, CA: National Marine Fisheries Service, Southwest Fisheries Science Center; 2007. p. 24.
30. Nowak RM. Walker's mammals of the world. 6th ed. Baltimore: Johns Hopkins University Press; 1999.
31. Salichos L, Rokas A. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature. 2013;497:327–31.
32. Shedlock AM, Takahashi K, Okada N. SINEs of speciation: tracking lineages with retroposons. Trends Ecol Evol. 2004;19:545–53.
33. Ray DA, Xing J, Salem A-H, Batzer MA. SINEs of a nearly perfect character. Syst Biol. 2006;55:928–35.
34. Suh A, Smeds L, Ellegren H. The dynamics of incomplete lineage sorting across the ancient adaptive radiation of Neoavian birds. PLoS Biol. 2015;13:e1002224.
35. McGowen MR, Spaulding M, Gatesy. Divergence date estimation and a comprehensive molecular tree of extant cetaceans. Mol Phylogenet Evol. 2009;53:891–906.
36. Martin AP, Palumbi SR. Body size, metabolic rate, generation time, and the molecular clock. Proc Natl Acad Sci U S A. 1993;90:4087–91.
37. Jackson J a, Baker CS, Vant M, Steel DJ, Medrano-González L, Palumbi SR. Big and slow: phylogenetic estimates of molecular evolution in baleen whales (suborder Mysticeti). Mol Biol Evol. 2009;26:2427–40.
38. Stewart C, Kural D, Strömberg MP, Walker J a, Konkel MK, Stütz AM, et al. A comprehensive map of mobile element insertion polymorphisms in humans. PLoS Genet. 2011;7:e1002236.
39. Arnold ML. Divergence with genetic exchange. New York: Oxford University Press; 2015. p. 272.
40. Suh A. The phylogenomic forest of bird trees contains a hard polytomy at the root of Neoaves. Zool Scr. 2016;45:50–62.
41. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.
42. Yim H-S, Cho YS, Guang X, Kang SG, Jeong J-Y, Cha S-S, et al. Minke whale genome and aquatic adaptation in cetaceans. Nat Genet. 2014;46:88–92.
43. DeWoody JA, Fernandez NB, Brüniche-Olsen A, Antonides JD, Doyle JM, San Miguel P, et al. Characterization of the gray whale *Eschrichtius robustus* genome and a genotyping array based on single-nucleotide polymorphisms in candidate genes. Biol Bull. 2017;232:186–97.
44. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
45. Li H, Durbin R. Fast and accurate long-read alignment with burrows-wheeler transform. Bioinformatics. 2010;26:589–95.
46. Shimamura M, Abe H, Nikaido M, Ohshima K, Okada N. Genealogy of families of SINEs in cetaceans and artiodactyls: the presence of a huge superfamily of tRNA(Glu)-derived families of SINEs. Mol Biol Evol. 1999;16:1046–60.
47. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005;110:462–7.
48. Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, et al. BEDOPS: high-performance genomic feature operations. Bioinformatics. 2012;28:1919–20.
49. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
50. Ortiz EM. vcf2phylip v1.5: convert a VCF matrix into several matrix formats for phylogenetic analysis. 2018. Available from: https://doi.org/10.5281/zenodo.1257058
51. Swofford D. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sunderland: Sinauer Associates; 2002.
52. Ronquist F, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, et al. Mrbayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 2012;61:539–42.
53. Bandelt HJ, Forster P, Rohl A. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol. 1999;16:37–48.
54. Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics. 2017;33:2938–40.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC

Supplementary material

Retrophylogenomics in rorquals indicate large ancestral population sizes and a rapid radiation

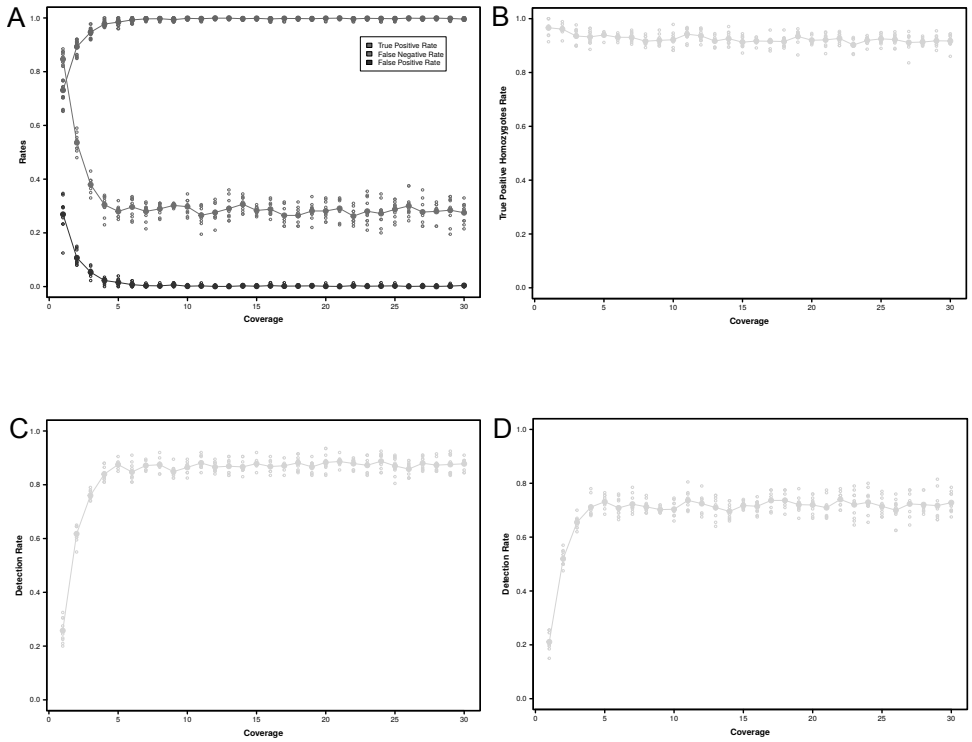
Fritjof Lammers, Moritz Blumer, Cornelia Rücklé, Maria A Nilsson

Supplementary Tables

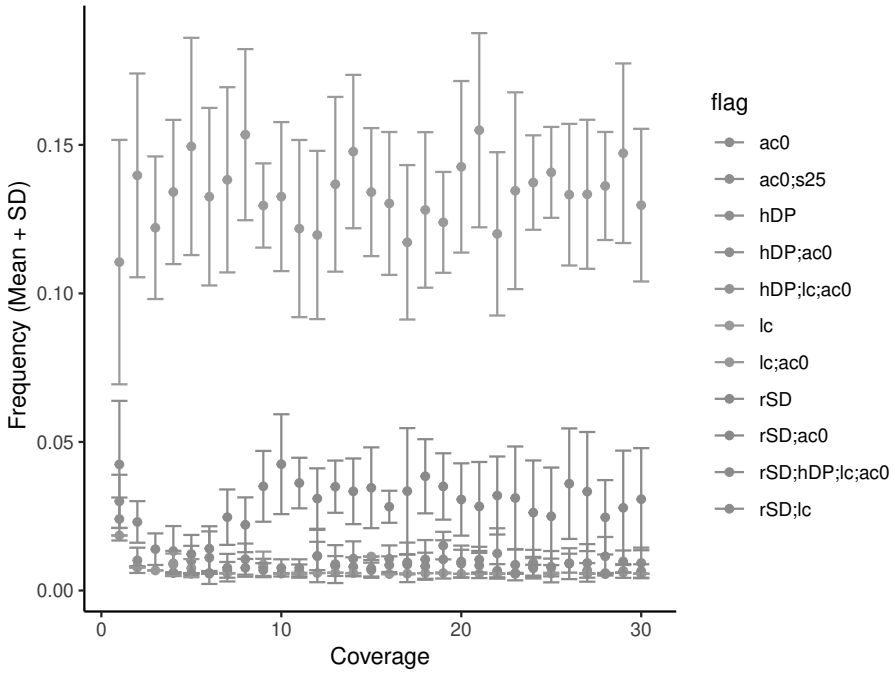
Supplementary Table 1. List of samples with accession numbers and sequencing properties

Species	ID	Accession-ID	Reference	Insert Size	Coverage
NA right whale	EgI0	SRR5665640	Arnason et al. 2018	470	10.69x
Sei whale A	Bbo01	SRR5665645	Arnason et al. 2018	482	10.36x
Sei whale B	Bbo02	SRR5665646	Arnason et al. 2018	482	10.27x
Blue whale	Bmu00	SRR5665644	Arnason et al. 2018	293	30.68x
Humpback whale	Mno00	SRR5665639	Arnason et al. 2018	470	27.88x
Fin whale	Bph03	SRR5665643	Arnason et al. 2018	294	13.99x
Minke whale	Bac00	SRR896642	Yim et. al 2014	472	7.73x
Gray whale (eastern) A	Ero01	SRR5665641	Arnason et al. 2018	284	17.60x
Gray whale (eastern) B	Ero02	SRR5665642	Arnason et al. 2018	284	22.29x
Gray whale (western) A	Ero03	SRR5495108	DeWoody et. al. 2017	497	29.24x
Gray whale (western) B	Ero04	SRR5495104	DeWoody et. al. 2017	471	26.65x

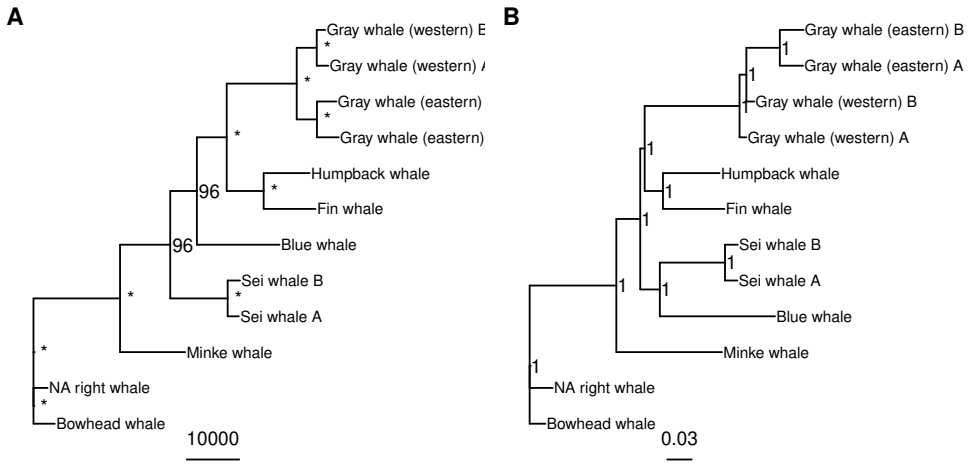
Supplementary Figures



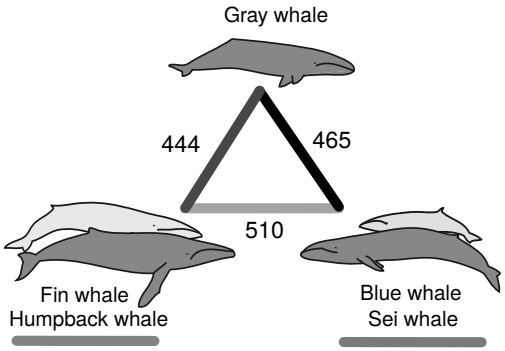
Supplementary Figure 1. Simulation results for CHR2 detection with MELT at varying depth of coverage using dataset specific parameters. A) Accuracy rates for detected CHR2 calls at varying coverages. B) Genotyping accuracy estimated from CHR2 insertions correctly called homozygous. C) MELT detection rate for the unfiltered and D) filtered datasets using MELTs internal filtering.



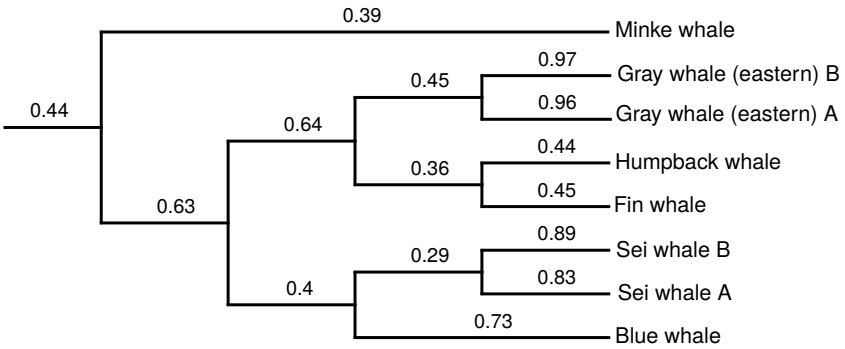
Supplementary Figure 2. Frequency of filters applied by MELT to exclude low-quality CHR2 calls. Most calls were excluded for proximity to low-complexity regions (lc). The other filter flags indicate sites without genotyped allele (ac0), sites that are not called in at least 25% of individuals (s25), sites with a high degree of discordant read pair evidence (hDP) and sites having unbalanced evidence on the 5' and 3' flanks of the insertion site (rSD). The description of filter flags are according to the MELT Documentation (http://melt.igs.umaryland.edu/manual.php).



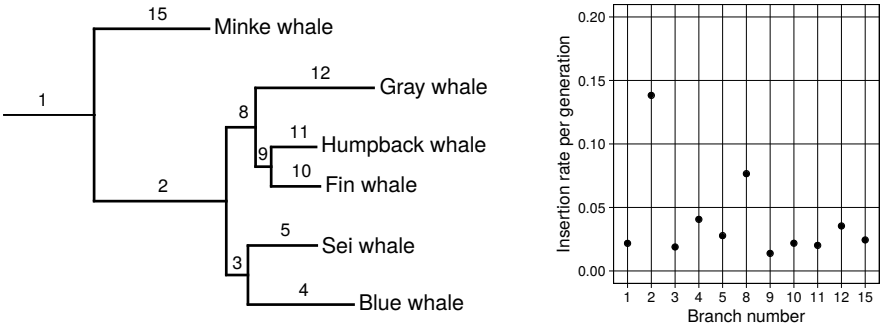
Supplementary Figure 3. Phylogenetic trees of baleen whales reconstructed with CHR2 insertions. A) Dollo-Parsimony tree reconstructed in PAUP*. Asteriks indicate 100 % bootstrap support (500 replicates), lower bootstrap support is given as numbers. B) Bayesian inference tree with posterior probability given for nodes.



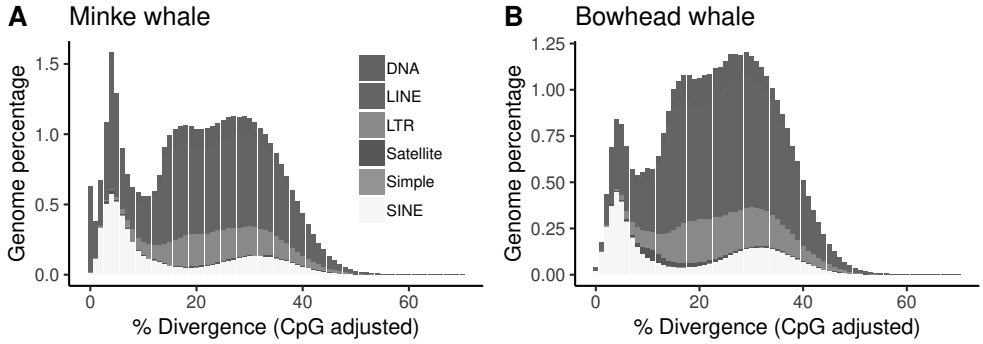
Supplementary Figure 4. Three alternative relationships in the rorqual radiation and the number of CHR2 insertion that support them. The KKSC test marginally rejects polytomy at p=0.02 (bifurcation test). The colors represent the edges in the network and signal quantification in Figure 2.



Supplementary Figure 5. Phylogenetic tree of rorquals with frequency of heterozygous insertions per branch. Heterzygous insertion frequency are calculated among extant species and mapped to the branches. Gray whale heterozygosity rates were calculated for each population (eastern and western Pacific).



Supplementary Figure 6. CHR2 insertion rates per generation. Divergence times are taken from reference no. 10. The average insertion rate of baleen whales was calculated to be 24.4 years.



Supplementary Figure 7. Repeatlandscapes of minke whale and bowhead whale based on available assemblies. The diagrams show the genome percent of the major repeat types depending on how much they have diverged from the consensus sequences.