Ancient and Modern Genomes Reveal Microsatellites Maintain a Dynamic Equilibrium Through Deep Time

McComish, Bennet J.; Charleston, Michael A.; Parks, Matthew; Baroni, Carlo; Salvatore, Maria Cristina; Li, Ruiqiang; Zhang, Guojie; Millar, Craig D.; Holland, Barbara R.; Lambert, David M.

Published in: Genome Biology and Evolution

DOI: 10.1093/gbe/evae017

Publication date: 2024

Document version: Publisher's PDF, also known as Version of record

Document license: CC BY-NC

Ancient and Modern Genomes Reveal Microsatellites Maintain a Dynamic Equilibrium Through Deep Time

Bennet J. McComish 1,2, Michael A. Charleston 1, Matthew Parks 3,4, Carlo Baroni 5,6, Maria Cristina Salvatore 5,6, Ruiqiang Li 7, Guojie Zhang 8,9, Craig D. Millar 10, Barbara R. Holland 1, and David M. Lambert 3,*

1School of Natural Sciences, University of Tasmania, Hobart, TAS 7001, Australia
2Menzies Institute for Medical Research, University of Tasmania, Hobart, TAS 7001, Australia
3Australian Research Centre for Human Evolution, Griffith University, Nathan, QLD 4111, Australia
4Department of Biology, University of Central Oklahoma, Edmond, OK 73034, USA
5Dipartimento di Scienze della Terra, University of Pisa, Pisa, Italy
6CNR-IGG, Institute of Geosciences and Earth Resources, Pisa, Italy
7Novogene Bioinformatics Technology Co. Ltd., Beijing 100083, China
8China National GeneBank, BGI-Shenzhen, Shenzhen 518083, China
9Department of Biology, Centre for Social Evolution, University of Copenhagen, Copenhagen DK-2100, Denmark
10School of Biological Sciences, University of Auckland, Auckland, New Zealand

*Corresponding author: E-mail: d.lambert@griffith.edu.au.
Accepted: January 23, 2024

Abstract

Microsatellites are widely used in population genetics, but their evolutionary dynamics remain poorly understood. It is unclear whether microsatellite loci drift in length over time. This is important because the mutation processes that underlie these important genetic markers are central to the evolutionary models that employ microsatellites. We identify more than 27 million microsatellites using a novel and unique dataset of modern and ancient Adélie penguin genomes along with data from 63 published chordate genomes. We investigate microsatellite evolutionary dynamics over 2 timescales: one based on Adélie penguin samples dating to ∼46.5 ka and the other dating to the diversification of chordates aged more than 500 Ma. We show that the process of microsatellite allele length evolution is at dynamic equilibrium; while there is length polymorphism among individuals, the length distribution for a given locus remains stable. Many microsatellites persist over very long timescales, particularly in exons and regulatory sequences. These often retain length variability, suggesting that they may play a role in maintaining phenotypic variation within populations.

Key words: microsatellite evolution, ancient DNA, Adélie penguin.

Significance

Microsatellites, or short tandem repeats, are widely used in population genetics, but their evolutionary dynamics remain poorly understood. We show that microsatellite loci remain stable yet variable for hundreds of millions of years, particularly those in exons and regulatory sequences. This leads us to recommend that population geneticists and ecologists use models of microsatellite evolution with stationary distributions, rather than those that allow allele lengths to drift upward indefinitely.
Introduction

Microsatellites or short tandem repeats, consisting of tandem repeats of motifs up to 6 bp in length, are prevalent in both prokaryotic and eukaryotic genomes. Some microsatellites have been shown to be functionally important (Kashi and King 2006; Mirkin 2007; Gemayel et al. 2010), but most are assumed to evolve neutrally, and for this reason, along with their abundance and high variability, they have been used extensively in population genetics studies (Schröter 2004; Voicu et al. 2021; Shi et al. 2023). While there have recently been important advances in our understanding of their genomic distribution (Srivastava et al. 2019) and evolutionary dynamics (Jonika et al. 2020; Verbiest et al. 2022), it remains unclear whether microsatellite loci are in dynamic equilibrium with respect to the length of alleles or whether alleles experience directional drift in length.

It is known that the processes governing microsatellite evolution can vary by genomic region. For instance, Zhang et al. (2006) investigated microsatellites in noncoding regions of the Arabidopsis genome; they found that the microsatellites conserved over evolutionary timescales were overrepresented in regulatory regions. Fujimori et al. (2003) found evidence for the involvement of microsatellites in gene regulation in plant genomes with a tendency for microsatellites to occur near the 5′ end of coding regions. They found very different compositions of microsatellites by genomic region in plants and animals, with animals having a far higher proportion of microsatellites in intronic regions relative to exons compared with plants. Understanding the different constraints on microsatellite evolution is important because the mutation processes that underlie these important genetic markers are central to the evolutionary models that employ microsatellites.

In this study, when describing microsatellites, we consider both the number of base pairs in the underlying motif (the period) and, for each allele, how many times the motif appears (the repeat number). We refer to microsatellites that contain only exact copies of the motif as pure. The total length of a microsatellite allele (in nucleotides) is the product of the period and repeat number. Repeat number is thought to change through a process of replication slippage (Levinson and Gutman 1987; Ellegren 2000), by which strands may transiently dissociate during DNA replication and then mispair with a different copy of the repeat, resulting in the insertion or deletion of one or more repeat units. Microsatellites are highly labile in evolutionary terms, with mutation rates resulting from replication slippage generally several orders of magnitude higher than for point mutation (Bhargava and Fuentes 2010).

Much still remains to be learned about the mutational processes involved in microsatellite evolution. The overall process can be thought of as a birth–death process (increase or decrease in length of a microsatellite by the birth or death of individual repeat units) embedded within a second birth–death process (microsatellite loci appear and disappear). Slippage during DNA replication is thought to be the main cause of changes in length, with mismatch repair reducing the mutation rate (Schröter 2000), but recombination may also play a role, and point mutations must be taken into account. Recent work by Murat et al. (2020) investigated the kinetics of DNA synthesis at microsatellites and found both slippage and point mutation to be associated with the secondary structures formed by some microsatellite sequences. The processes by which new microsatellites appear and eventually degenerate and disappear are particularly poorly understood (Kelkar et al. 2011) and may account for shifts in the relative abundance of microsatellite types in closely related species, which cannot be explained by strand slippage alone.

Existing models of microsatellite evolution are highly simplified and take into account only changes in length (and occasionally purity) in existing microsatellites, ignoring the processes of “birth” and “death” by which microsatellite loci appear and disappear (and perhaps reappear; Buschiazzo and Gemmell 2006). Some of these models have been designed so that they have a stationary distribution (e.g. those of Kruglyak et al. 1998; Calabrese et al. 2001; Amos et al. 2015), but it is not clear whether this is biologically realistic. It may be that an individual microsatellite locus is never at equilibrium, tending instead to increase in length throughout its life, but that the birth–death process causes the genome-wide distribution of allele lengths at all microsatellite loci to be at equilibrium.

A key open question is thus whether the alleles at a microsatellite locus increase or decrease in average length over time or whether each locus is maintained at an equilibrium length. Some pedigree studies have shown a bias in favor of the gain of repeats (Weber and Wong 1993), suggesting that microsatellites should rapidly increase in size (Rose and Falush 1998). Based on these observations, it has been suggested that microsatellites increase in length until the accumulation of point mutations hinders slippage and ultimately leads to the degeneration of the microsatellite locus (Kruglyak et al. 1998; Buschiazzo and Gemmell 2006). Other studies have found that slippage has a length-dependent bias (Xu et al. 2000; Huang et al. 2002; Sun et al. 2012), supporting earlier suggestions that constraints exist on repeat number at microsatellite loci (Garza et al. 1995). Amos et al. (2015) proposed a model consistent with these observations, in which interallelic interactions in heterozygous individuals may drive the process whereby longer-than-average alleles tend to get shorter and shorter-than-average alleles tend to get longer (which they call the centrally directed mutation model).

In this study, we make use of exceptionally well-preserved ancient DNA from a unique set of Adélie penguin...
samples reported here for the first time and genotype 177,974 microsatellites of periods 2 to 6 in both modern and ancient genomes, including some dating to ~46.5 ka. In addition, we are able to time the evolutionary origin of many of these loci by aligning them with more than 27 million microsatellite loci from a large set of published chordate genomes and mapping them onto an appropriate phylogeny (Jarvis et al. 2014). Our data include microsatellites that date to the diversification of chordates aged more than 500 Ma. We show that allele lengths at microsatellite loci are in dynamic equilibrium, and these have remained stable over hundreds of millions of years and through many speciation events. While there is length polymorphism among individuals, the overall length distribution for a given locus does not change appreciably over time. We show that microsatellites can persist over very long timescales, particularly those in exons and regulatory sequences, while retaining length variability. This suggests that microsatellites may play a role in maintaining phenotypic variation within populations.

Results

Microsatellite Dynamics in Adélie Penguin Samples

Genomes obtained from ancient biological remains allow us to observe changes in sequence variation that cannot be observed using only contemporary sequences. In this study, we have used whole-genome sequence data of ancient Adélie penguin remains from 23 individuals dated up to 46,587 years old, as well as from 26 modern individuals, to genotype 177,974 microsatellite loci of periods 2 to 6 identified in an Adélie penguin reference genome. Most loci are close to the minimum length detectable for each period (especially in the case of pure loci), with very small numbers of longer loci up to thousands of base pairs in length. The length distributions of these microsatellite loci are shown in supplementary fig. S1, Supplementary Material online. We determined the genotype of these loci in each of the ancient and modern Adélie samples, and allele length distributions for each sample are shown in supplementary fig. S2, Supplementary Material online. These genotype data enable us to obtain length distributions for microsatellites at different time points and hence to test for any evidence of directional drift in microsatellite length.

To test whether microsatellite allele length is stationary or whether the average allele lengths of individual loci increase over time, we used BayesFactor (Morey and Rouder 2015) to compare generalized linear mixed models in which allele length is treated as dependent on different combinations of possible explanatory variables. The explanatory variables considered were as follows: the motif of the allele, the surrounding sequence type (exon, intron, regulatory, or intergenic), and sample age. We also tested for any interactions between the surrounding sequence type and sample age. In addition to these fixed effects, which are assumed to be the same for all genomes, we also treated the locus as a random effect; this is equivalent to allowing a different intercept in the regression model for each locus in the genome. Impurity affects the length at which microsatellites can be detected, so models were fit separately for pure and impure microsatellites. Similarly, different alignment score thresholds were used to detect loci of different periods, so that their allele lengths cannot be compared directly, and models were, therefore, fit separately for each period. Because the large numbers of loci made testing the full dataset impractical, we tested 5 random samples of 2,000 loci for each purity/period combination.

Bayes factors (BFs) for all models tested are given in supplementary table S1, Supplementary Material online, and posterior estimates of effect sizes are given in supplementary table S2, Supplementary Material online. For both pure and impure microsatellites of each period, the best-supported model for each sample is that in which allele length is dependent on the motif and locus and, in most cases, the surrounding sequence type. The only exception to this pattern is for impure loci of period 2, for which the best-supported model in 4 out of 5 samples is that in which allele length depends on locus and location, independent of the motif. Our data provide evidence against models in which length depends on sample age, being at least 5 times less likely to be observed under these models than under the best-supported model in all cases. Since length does not depend on sample age in the best-supported models, we infer that the process of expansion and contraction of microsatellite alleles is effectively stationary, or nearly stationary, over a timescale of tens of thousands of years.

Microsatellite Locus Age Inference

To investigate microsatellite dynamics over a much longer timescale and across a broad range of species, we used whole-genome alignments of 48 avian species from Zhang et al. (2014) along with the genomes of 15 non-avian vertebrate species that span the chordate tree. We identified a total of over 27 million microsatellites in the 63 genomes, and a breakdown of the number of loci of each period detected in each genome is given in supplementary table S3, Supplementary Material online. For each of these species, we used a whole-genome alignment to chicken to generate a standard set of coordinates for all microsatellites present in the alignment. The number of microsatellite loci in any species that can be aligned to the chicken genome, and the overall number of bases aligned to the chicken genome, are negatively correlated.

References

Jarvis et al. (2014)

To cite this article, you could use the following format:


with the time since the most recent common ancestor of that species and chicken (see supplementary fig. S3, Supplementary Material online). We were able to map ~5.4 million microsatellites across the 63 species to almost 2.9 million loci in the chicken genome. Of these, almost 2.2 million microsatellite loci were found in only a single species, and 680,804 loci had microsatellites conserved across 2 or more species. The exact numbers of microsatellites detected and aligned are given in supplementary table S4, Supplementary Material online.

We used the dated avian whole-genome phylogeny published by Jarvis et al. (2014), to which we added 15 non-avian species with estimated divergence times taken from the Timetree of Life (www.timetree.org, adjusted times; Hedges et al. 2015). To infer gains and losses of microsatellite loci in different lineages, we carried out ancestral state reconstruction on a subtree whose topology is relatively uncontroversial, agreeing with the trees published by Jarvis et al. (2014) and Prum et al. (2015), and on which we expect incomplete lineage sorting events to be rare (Suh et al. 2015). This allows us to infer the edge on which any locus present in Adélie penguin was gained and hence to estimate the ages of these loci. Distributions of estimated ages for loci in intergenic, intronic, exonic, and regulatory sequences are shown in supplementary fig. S4, Supplementary Material online. Supplementary fig. S5, Supplementary Material online shows the numbers of inferred gains and losses of microsatellites on each edge of the subtree, scaled according to both the length of the edge and the amount of sequence that can be aligned to the chicken genome. The total number of extant microsatellite loci whose origins were inferred to predate selected ancestral nodes is given in supplementary table S5, Supplementary Material online.

The relative densities of microsatellite loci (including both pure and impure microsatellites) in different types of sequences for different age brackets are shown in Fig. 1. While the older age brackets contain fewer loci overall, those loci are much more likely to be found in regulatory or coding sequences. The percentage of loci found in regulatory or coding sequence for each bracket is shown in supplementary table S6, Supplementary Material online. Microsatellite loci in regulatory or coding sequences thus appear to be conserved over longer periods on average than those in intergenic sequences or introns. This suggests that they are maintained by selection, be it directly for the locus since the latter is confounded with the locus age estimate. In these models, we treated the sample as a random effect rather than the locus since the latter is confounded with the locus age estimate. This allowed us to test the models on all loci without having to sample. BFs for all the models tested are given in supplementary table S8, Supplementary Material online. For all subsets of the data comprising pure and impure microsatellites of each period, the best-fitting model for allele length is dependent on sample, motif, surrounding sequence type, locus age, and an interaction between surrounding sequence type and locus age. In all cases, the data provide very strong evidence for this model, being more likely under this model than under any other by a factor of at least $10^{12}$. We sampled from the posterior distribution of the full model for each subset to obtain posterior estimates of effect sizes, and these are shown in supplementary table S9, Supplementary Material online. The effect of locus age is shown separately in Table 1. The effect sizes are very small (on the order of one nucleotide per hundred million years). For loci of periods 2 and 3, we also tested interactions between motif and locus age and between motif and surrounding sequence type for subsets of our data and found strong evidence for these interactions. We were unable to test these interactions for loci of longer periods because of the rapid increase in numbers of motifs as the period increases.

Distributions of allele lengths for loci of different ages in different types of surrounding sequences are shown in Fig. 2. In agreement with the results of the linear mixed model, a very slow increase in mean allele length over time can be seen for microsatellites in intronic and intergenic sequences. Overall, pure dinucleotide and tetranucleotide loci in protein-coding sequences have significantly shorter mean allele lengths than those in noncoding sequences, while impure trinucleotide and hexanucleotide microsatellites in protein-coding sequences have significantly longer mean allele lengths than those in noncoding sequences (see Table 2). It is likely that selection against frameshift mutations in coding sequence limits microsatellite expansion when the period is not a multiple of 3 (Metzgar et al. 2000).

**Discussion**

To summarize our results, for the microsatellites investigated in this study, the mean allele length at any given microsatellite locus changes very little on scales ranging from a few thousand to hundreds of millions of years, with estimated effect sizes on the order of one nucleotide
per hundred million years. For microsatellite loci of periods 2 to 6 that can be mapped onto the avian phylogeny, there is a gradual increase in allele length variation over time, as can be seen in Fig. 2. This suggests that the replication slippage process that generates length polymorphism is in a dynamic equilibrium, such that increases and decreases in length remain approximately balanced. These results are consistent with the findings of Sun et al. (2012) that longer alleles tend to decrease in length and shorter alleles tend to increase. We recommend that population geneticists and ecologists use models of microsatellite evolution that have stationary distributions, such as those proposed by Kruglyak et al. (1998), Calabrese et al. (2001), or Amos et al. (2015), rather than those such as the stepwise mutation model (Ohta and Kimura 1973) that allow allele lengths to drift upwards indefinitely.

We have also shown that microsatellites can persist, and remain variable, over very long periods of evolutionary time, with 257 extant microsatellite loci dating from before the origin of chordates and 3,938 predating the divergence of mammals and reptiles. These numbers are underestimates, as they exclude any loci that have been lost in birds or that could not be aligned to the chicken reference genome we used. Although we observe a slight decrease in
heterozygosity with locus age (supplementary fig. S6, Supplementary Material online), we nevertheless observe multiple alleles in the Adélie samples for many ancient loci (supplementary table S7, Supplementary Material online). The microsatellite loci that persist over very long periods are more often found in coding sequences and in regulatory regions. A disproportionate number of these variable ancient loci are trimer repeats located in protein-coding genes, which must code for a homopolymer run of amino acids.

These trimer repeats in coding sequences make up only 0.55% of all loci that are variable in our Adélie samples, but 5.67% of variable loci that predate the divergence of extant birds and 9.86% of those that predate the divergence of mammals and reptiles. It seems likely that selection is acting to maintain variability at these loci, which could act as mediators of rapid phenotypic change (Gemayel et al. 2010).

A limitation of using short-read data is that longer alleles are effectively censored from our data; however, as can be seen in supplementary fig. S1, Supplementary Material online, the overwhelming majority of loci are much shorter (in the reference genome) than the read length. In addition, the reads from ancient Adélie samples are shorter than those from modern samples. This means that longer alleles are less likely to be genotyped in the ancient samples, and therefore, we would expect this to give a signal for increasing allele length over time. However, we do not observe any such signal despite this potential bias, presumably because any such signal is obscured by the much larger number of shorter loci. With the increasing availability of long-read sequencing and as methods for genotyping microsatellites in long-read data are developed, it may now be feasible to verify our results in a more complete dataset.

### Materials and Methods

#### Contemporary Adélie Penguin Samples

Blood samples from Adélie penguins were collected from individuals at active breeding colonies, using methods as described in Millar et al. (2008), in 6 locations around Antarctica: Torgerson Island (AP samples), the Mawson region (B samples), Cape Adare, Cape Bird, Coulman Island, Fieg. 2. — Distributions of allele lengths for loci of different ages in different types of surrounding sequences. Distributions of mean allele lengths (in nucleotides) of pure and impure microsatellite loci present in Adélie penguin and conserved across 6 age brackets for loci with periods 2 to 6 in intergenic, intron, exon, and regulatory sequences. The 6 age brackets in each cluster correspond to loci that arose most recently on the branch leading to Adélie penguin; on the branch leading to penguins; within Neoaves or on the branch leading to Neoaves; on the branch leading to Neognathae; on the branch leading to birds; outside Sauria (note that no pure exonic microsatellites of period 5 are inferred to have arisen outside Sauria). Each box extends from the lower to upper quartiles of the length distribution, and the interior line indicates the median. The whiskers extend to the most extreme points within 1.5× interquartile range of the quartiles.
Table 2
The mean and standard error of allele lengths at microsatellite loci in different types of surrounding sequences

<table>
<thead>
<tr>
<th>Period</th>
<th>Intergenic</th>
<th>Intron</th>
<th>Exon</th>
<th>Regulatory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pure</td>
<td>13.06 (0.0027)</td>
<td>12.93 (0.0039)</td>
<td>11.79 (0.0121)</td>
<td>12.79 (0.0100)</td>
</tr>
<tr>
<td>Impure</td>
<td>20.53 (0.0080)</td>
<td>20.11 (0.0129)</td>
<td>19.66 (0.0718)</td>
<td>21.03 (0.0303)</td>
</tr>
<tr>
<td>3</td>
<td>15.99 (0.0048)</td>
<td>15.70 (0.0071)</td>
<td>16.06 (0.0154)</td>
<td>16.03 (0.0158)</td>
</tr>
<tr>
<td></td>
<td>24.02 (0.0162)</td>
<td>23.68 (0.0302)</td>
<td>26.73 (0.0521)</td>
<td>22.79 (0.0402)</td>
</tr>
<tr>
<td>4</td>
<td>16.40 (0.0043)</td>
<td>16.03 (0.0058)</td>
<td>14.83 (0.0258)</td>
<td>15.61 (0.0103)</td>
</tr>
<tr>
<td></td>
<td>24.59 (0.0111)</td>
<td>24.12 (0.0174)</td>
<td>24.45 (0.1324)</td>
<td>23.15 (0.0360)</td>
</tr>
<tr>
<td>5</td>
<td>18.93 (0.0091)</td>
<td>18.35 (0.0127)</td>
<td>17.35 (0.0468)</td>
<td>17.70 (0.0226)</td>
</tr>
<tr>
<td></td>
<td>27.99 (0.0156)</td>
<td>26.84 (0.0244)</td>
<td>24.44 (0.1561)</td>
<td>25.86 (0.0524)</td>
</tr>
<tr>
<td>6</td>
<td>18.44 (0.0086)</td>
<td>18.02 (0.0124)</td>
<td>17.78 (0.0137)</td>
<td>17.95 (0.0280)</td>
</tr>
<tr>
<td></td>
<td>27.73 (0.0187)</td>
<td>26.74 (0.0274)</td>
<td>29.71 (0.0521)</td>
<td>26.75 (0.0870)</td>
</tr>
</tbody>
</table>

and Inexpressible Island. Collection and sequencing information is given in supplementary table S10, Supplementary Material online.

Ancient Adélie Penguin Samples

Sub-fossil bones were collected in abandoned nests discovered along coastal ice-free areas both in the vicinity of presently occupied colonies and in relict colonies discovered in sites where penguins do not breed at present (Baroni and Orombelli 1994; Baroni 2013; Lorenzini et al. 2014; supplementary table S11, Supplementary Material online). Ornithogenic soils were stratigraphically excavated to find penguin bones and other remains as described previously (Lambert et al. 2002; Ritchie et al. 2004).

Radiocarbon accelerator mass spectrometry (AMS) dates were supplied by NOSAMS, Woods Hole Oceanographic Institute, the New Zealand Institute of Geological and Nuclear Sciences, Lower Hutt, New Zealand, and Institut fôr Fysik og Astronomi, Aarhus Universitet, Denmark. Radiocarbon dates were calibrated with CALIB 7.1 (http://calib.qub.ac.uk/calib/; Reimer et al. 2013) using the Marine Reservoir Correction Database 2013 and applying a delta-R of 791 ± 121 (Hall et al. 2010). Mean ages and 2 delta standard deviation values were considered.

Modern DNA Extraction

For the 26 modern Adélie penguin samples, genomic libraries were prepared by first extracting DNA from Seutin-preserved blood or ethanol-preserved soft tissue samples. DNA was then purified using Qiagen MinElute or PCR Purification columns according to the manufacturer’s protocol (Qiagen, Valencia, CA, USA) and eluted in 100-µL UltraPure water (Life Technologies, Grand Island, NY, USA).

Ancient DNA Extraction

All laboratory work with ancient Adélie penguin samples prior to polymerase chain reaction (PCR) amplification of genomic libraries (see below) was carried out in a physically isolated laboratory used only for ancient DNA work, following strict guidelines to minimize external contamination. Designated blank samples consisting originally of 200 µL of digestion buffer were carried sequentially through all DNA extraction and library building procedures at a minimum ratio of 1 blank for every 8 samples.

DNA was extracted from ancient bone or muscle tissue samples by first digesting ∼0.1 g bone/tissue shavings in 200-µL digestion buffer (consisting of 180 µL of 0.5 M EDTA, 10 µL of 10% N-lauryl sarcosine, and 10 µL of 20 mg/mL proteinase K) for 12–18 h at 55 °C with rotational mixing (∼10 rpm). This was followed by 2–5 rounds of organic extraction with 1–1.5 mL of ultra-pure buffer-saturated phenol and 1 round of extraction with 1–1.5 mL of chloroform (Sigma-Aldrich, St. Louis, MO, USA). Extracts were purified with Qiagen MinElute or PCR Purification columns using high-concentration buffer PB or PE (10:1 buffer: sample volume ratio) to improve retention of small fragments and 2× spin-through centrifugation for sample application and elution stages to further maximize yield. Final elutions were completed in a volume of 22-µL UltraPure water or New England Biolabs (NEB) elution buffer (EB).

DNA Library Construction

Purified extracted DNA of modern samples was quantified with a Qubit 2.0 Fluorometer and dsDNA HS Assay kit (Life Technologies), and ∼0.5–1.5 µg of DNA was sheared with a Covaris sonicator (Covaris, Woburn, MA, USA) to an average size of 300–600 bp. Sheared extracts were adapter ligated and enriched using the standard NEBNext or NEBNext Ultra protocol (catalog #E6040 and #E7370) and NEBNext multiplex Illumina primers (catalog #E7335; NEB, Ipswich, MA, USA) in ½-size recommended reaction volumes for end-preparation, adapter ligation, and enrichment reactions. Enrichments were performed under recommended cycling conditions, with 10–14 cycles of enrichment for each sample and using Phusion High-Fidelity Master Mix (NEB catalog #M0531). Samples were submitted for 101-bp paired-end sequencing on an Illumina HiSeq2000 at BGI-Hong Kong, using 1 or ∼1.33 lanes for each sample.

Genomic libraries for ancient samples were built following 2 strategies. Library building for all Holocene samples and initial attempts for 2 late Pleistocene samples (CB070121.08 and CB070121.16) were completed based on Meyer and Kircher (2010) with minor adjustments. Based on low endogenous yields for the 2 late Pleistocene
samples, a second attempt at library building was made for all 3 late Pleistocene samples (CB070121.08, CB070121.13, and CB070121.16) following the NEBNext Ultra protocol and using ½-size reactions for end-preparation, adapter ligation, and enrichment reactions. For all ancient samples, enrichment reactions were completed by mixing ~11.5 µL of the heat-inactivated adapter ligation reaction, 0.5 µL each of 25 µM NEBNext index and universal primers, and 12.5 µL of 2x Phusion Hi-Fidelity Master Mix. Enrichment reactions were carried out under recommended cycling conditions, with 12–22 cycles of enrichment for each sample.

Finished ancient libraries were purified using Axygen MAG-PCR SPRI beads (Corning Life Sciences, Tewksbury, MA, USA) at a ratio of 0.7 to 1:1:1 Axygen:sample volume to minimize the concentration of potential adapter dimers (Quail et al. 2009) and quantified with a Qubit 2.0 Fluorometer. Libraries were submitted for 101-bp single-end sequencing on an Illumina HiSeq2000 to either BGI-Hong Kong or the National High-Throughput Sequencing Center (University of Denmark, http://seqcenter.ku.dk/), using between 2 and 10.5 lanes of sequencing for each sample with resultant genome-wide average sequencing depths of ~22x and 8x for modern and ancient samples, respectively (supplementary tables S10 and S11, Supplementary Material online).

Alignment

For all sequence pools, adapter sequences were trimmed from reads using Cutadapt (Martin 2011) v. 1.1 under default parameters. Low-quality reads were filtered with Trimmomatic (Bolger et al. 2014) v. 0.22, with a minimum trailing and leading quality of 20, an average quality over 20-bp sliding windows of 20, and minimum lengths of 80 bp for modern reads and 30 bp for ancient reads. Trimmed and filtered Illumina reads for each Adélie penguin sample were mapped to the Adélie reference genome (Zhang et al. 2014) using Bowtie2 (Langmead and Salzberg 2012) with the “--very-sensitive” preset option.

Genomes

In this study, we use the 48 avian genomes reported by Jarvis et al. (2014). We also use the pairwise alignments to the chicken genome that were used by Jarvis et al. (2014) in generating their whole-genome multiple alignments. This consists of a set of pairwise alignments for each species with each individual chromosome of the chicken genome as reference.

In addition, we use genomes of the 15 non-avian species for which whole-genome alignments to the chicken galGal3 assembly are available from the UCSC genome browser. These are as follows: human (hg19), chimpanzee (panTro3), orangutan (ponAbe2), mouse (mm9), rat (rn4), guinea pig (cavPor3), horse (equCab2), opossum (monDom5), platypus (ornAna1), lizard (anoCar2), frog (xeNTro3), zebrafish (danRer4), fugu (fr2), lamprey (petMar1), and lancelet (braFlo1). All genomes used are listed in supplementary table S12, Supplementary Material online.

Microsatellite Detection

Microsatellite loci were identified in all 63 genomes using Tandem Repeats Finder (TRF; Benson 1999) with the following parameters: match weight 2; mismatch weight 7; indel weight 7; matching probability 80; indel probability 10; minimum alignment score 18; and maximum period size 6. The results were then filtered using the alignment score thresholds shown in supplementary table S13, Supplementary Material online, taken from Willems et al. (2014). This gave us 5 sets of microsatellites for each species: dimer, trimer, tetramer, pentamer, and hexamer repeats, with their respective score thresholds. Mononucleotide repeats were not included for the reasons outlined in Willems et al. (2014), i.e. because they are more prone to PCR stutter artifacts, particularly with low sequencing coverage, and this would likely be exacerbated by the enrichment reactions used for our ancient samples.

Microsatellite loci were compared against the annotations for all the avian genomes, to determine which loci fall within protein-coding sequences or introns. Putative regulatory regions were identified by extracting the set of conserved nonexonic elements identified in the chicken genome by Lowe et al. (2015) from each of the avian genome alignments. All remaining sequences were assumed to be intergenic.

Microsatellites identified in the Adélie penguin reference genome using TRF were genotyped in the Adélie penguin samples using RepeatSeq (Highnam et al. 2013) (which requires a list of preidentified loci and sequence reads as input), and the output was formatted as tables for analysis in R (R Development Core Team 2011). Tables of genotype calls were imported into R and summary statistics calculated for each locus, including the mode, mean, and standard deviation of the allele lengths observed in the samples, and the number of alleles observed. These were combined with the TRF output containing the motif, purity, and nucleotide composition of the locus in the reference genome.

Homology Matching

First, for each species and period, we coded the microsatellite loci detected above as features in a general feature format file. Next, we used MapFilter (Dutheil et al. 2014) to extract these features from the pairwise alignment between the species in question and each chicken chromosome and output the coordinates that each feature aligns to in the chicken genome. A custom R script was used to produce a table matching each set of chicken coordinates...
to the corresponding microsatellite locus. Motifs were standardized by calculating the lexicographically minimal rotation to allow for loci to begin at different positions within a repeat unit (e.g. the motif TGA was standardized as ATG).

For each chicken chromosome and period, we combined the motif tables for all 63 species and used a custom Java program to assign similarity scores to pairs of loci based on the distance between them (in terms of chicken coordinates) and the similarity of their motifs. Loci were scored if they were no more than 60 bp apart, and their motifs differed by no more than one substitution. Testing different values of the length threshold showed that larger values did not increase the numbers of homologous loci detected. In addition, we manually checked a small sample of loci to verify that the loci detected were indeed homologous. The similarity was calculated as

\[
\text{sim} = \frac{\rho}{1 + d}
\]

where \(\rho\) is the proportion of sites in the motif that are identical, and \(d\) is the distance in base pairs between the loci (0, if the loci overlap). We then used the Markov Cluster Algorithm (van Dongen 2008) with the \(-abc\) input option and default settings to identify clusters of putatively homologous loci (across all 63 genomes). These clusters were converted into a binary present/absence matrix with the 48 species as columns and locations as rows, where a one represents the presence of a microsatellite for a given combination of species and location and a zero indicates that a species does not have a microsatellite in that location.

To avoid any false negatives where a given region is not represented in the alignment for some species, we checked the local region of the alignment for any species missing (“?,” as opposed to “0” for absent) in the presence/absence matrix if the region was not covered in the alignment.

**Ancestral State Reconstruction**

We used the R package phangorn (Schliep 2011) to perform ancestral state reconstruction on the timetree reported in Jarvis et al. (2014) using our presence/absence matrices. The maximum likelihood reconstructions available do not allow nonreversible models (i.e. the rates of gains and losses are assumed to be equal), so we used the “ACCTRAN” parsimony method. Numbers of gains and losses of homologous microsatellites inferred for each edge were then counted, ignoring any changes from a known state to an unknown one.

We also calculated the numbers of microsatellite losses required under a Dollo process, where any microsatellite locus only ever arises once but may be lost in multiple lineages. However, the results were not appreciably different from those obtained under parsimony.

**Adélie Locus Age Determination**

Minimum ages were calculated for loci present in the Adélie penguin genome by using the ancestral state reconstruction results to identify the most recent gain of the locus on the path from the root to Adélie. This allows for loci to be gained independently in different lineages or lost and regained. These locus ages were combined with the genotype statistics calculated above, allowing us to examine the relationship among locus age, length, purity, and surrounding sequence type.

**Model Fitting**

The “generalTestBF” function of the R package BayesFactor (Morey and Rouder 2015) was used to fit generalized linear mixed models to the ancient and modern Adélie genotype data. A BF is a measure that quantifies the evidence for a hypothesis compared with an alternative hypothesis given the data. The following thresholds have been suggested to quantify the evidence for one hypothesis over another as reported by BFs: BF < 3, insignificant; BF 3–20, positive; BF 20–150, strong; and BF > 150, very strong (Kass and Raftery 1995).

We tested the dependence of microsatellite allele length on sample age, motif, locus, surrounding sequence type, and an interaction between sample age and surrounding sequence type for all loci genotyped in Adélie. Because of the large number of loci, we tested 5 random samples of 2,000 loci for pure and impure microsatellite loci of each period. For those loci for which we were able to estimate the age (i.e. those that were alignable to the chicken genome), we tested the dependence of allele length on estimated locus age, motif, sample, surrounding sequence type, and an interaction between locus age and surrounding sequence type. Locus and sample were treated as random effects, and all other variables were treated as fixed effects. Sample age and locus age variables were centered. Models were fit separately for pure and impure microsatellite loci of each period (2 to 6). To obtain estimates of effect sizes, we used the “posterior” function of BayesFactor to generate samples from the posterior distributions of the full models. We also tested interactions between motif and locus age and between motif and surrounding sequence type.

Our workflow for detecting homologous microsatellite loci and estimating their ages, starting from genome sequences and pairwise alignments, is given in supplementary fig. S7, Supplementary Material online.

**Supplementary Material**

Supplementary material is available at Genome Biology and Evolution online.

**Acknowledgments**

We thank John Macdonald and Peter Ritchie for assistance with the collection of contemporary Adélie penguin samples.
Author Contributions

Funding
This work was supported by a Human Frontiers Science Program grant (RGP0036) and an Australian Research Council grant (2157200). Preliminary studies were funded by the Australia–India Strategic Research Fund to D.M.L. In addition, we thank Griffith University and the University of Tasmania for support and the BGI for sequencing of contemporary Adélie penguins (grant D0170101313) and the Copenhagen DNA Sequencing Facility for ancient DNA sequencing. We are grateful to the Italian National Program on Antarctic Research (PNRA-4.2/2004) and Antarctica New Zealand for providing support for the Antarctic fieldwork.

Conflict of Interest
All authors declare that they have no competing interests.

Data Availability
The datasets generated and analyzed during the current study, and the code used for analysis, are available in the Dryad repository, https://doi.org/10.5061/dryad.7gt3rg2. The Adélie penguin sequence read data have been deposited with links to BioProject accession number PRJNA210803 in the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/).

Literature Cited
Koehler MR, Martin JA, Davis AF, Kinghorn JR. The Adélie penguin sequence read data have been deposited to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/).