MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2016) **16**, 314–324 doi: 10.1111/1755-0998.12443

The *de novo* genome assembly and annotation of a female domestic dromedary of North African origin

ROBERT R. FITAK,*[1] ELMIRA MOHANDESAN,* JUKKA CORANDER† and PAMELA A. BURGER*

**Institut für Populationsgenetik, Vetmeduni Vienna, Veterinärplatz 1, Vienna 1210, Austria, †Department of Mathematics and Statistics, University of Helsinki, Helsinki, FIN-0014, Finland*

Abstract

The single-humped dromedary (*Camelus dromedarius*) is the most numerous and widespread of domestic camel species and is a significant source of meat, milk, wool, transportation and sport for millions of people. Dromedaries are particularly well adapted to hot, desert conditions and harbour a variety of biological and physiological characteristics with evolutionary, economic and medical importance. To understand the genetic basis of these traits, an extensive resource of genomic variation is required. In this study, we assembled at 65× coverage, a 2.06 Gb draft genome of a female dromedary whose ancestry can be traced to an isolated population from the Canary Islands. We annotated 21 167 protein-coding genes and estimated ~33.7% of the genome to be repetitive. A comparison with the recently published draft genome of an Arabian dromedary resulted in 1.91 Gb of aligned sequence with a divergence of 0.095%. An evaluation of our genome with the reference revealed that our assembly contains more error-free bases (91.2%) and fewer scaffolding errors. We identified ~1.4 million single-nucleotide polymorphisms with a mean density of 0.71×10^{-3} per base. An analysis of demographic history indicated that changes in effective population size corresponded with recent glacial epochs. Our *de novo* assembly provides a useful resource of genomic variation for future studies of the camel's adaptations to arid environments and economically important traits. Furthermore, these results suggest that draft genome assemblies constructed with only two differently sized sequencing libraries can be comparable to those sequenced using additional library sizes, highlighting that additional resources might be better placed in technologies alternative to short-read sequencing to physically anchor scaffolds to genome maps.

Keywords: adaptation, *Camelus dromedarius*, demography, domestication, next-generation sequencing

Received 2 February 2015; revision received 22 June 2015; accepted 25 June 2015

Introduction

The dromedary (*Camelus dromedarius*) is the most common of all *Camelus* species and is easily distinguished from its congeners, the Bactrian (*Camelus bactrianus*) and wild (*Camelus ferus*) camels, by the presence of a single hump. Dromedaries are widespread throughout northern and eastern Africa, the Arabian Peninsula and southwest Asia, and a large feral population exists in Australia (Köhler-Rollefson 1991; Spencer & Woolnough 2010). Throughout their range, dromedaries are bred for a multitude of purposes including meat, milk production, transportation, wool and sport (Bulliet 1990; Grigson 2012). Archaeozoological evidence suggests that the domestication of dromedaries took place between 3000 and 4000 years ago in the east coast of the Arabian Peninsula (Uerpmann & Uerpmann 2002). Unlike many other domestic livestock, the wild ancestor of dromedaries is extinct, and despite the examination of ancient wild dromedary remains, a formal taxonomic description of the extinct species has not been made.

In addition to the economic importance, dromedaries harbour an assortment of biological and physiological traits specifically adapted to extreme heat and harsh, desert conditions. For example, dromedaries do not begin to sweat until body temperatures reach as high as 42 °C, can tolerate fluctuating body temperatures as much as 6 °C and withstand water loss >30% of their body mass (see Köhler-Rollefson 1991 for a review). Furthermore, studies have uncovered several camel products with applications in human medicine, including unique immunoglobulin molecules that are useful in nanobody technology (Muyldermans *et al.* 2009) and milk that may contain beneficial properties for the treatment of diabetes (Agrawal *et al.* 2011).

Correspondence: Robert R. Fitak, Fax: 919-660-7293;
E-mail: robert.fitak@duke.edu

[1]Present address: Department of Biology, Duke University, Durham, NC 27708, USA.

As a result of increased economic, medical and evolutionary value of camels, understanding the genetic basis of these and other relevant traits is necessary. However, unlike many other livestock species (e.g. cow, horse, pig), genetic and genomic resources for camels, especially dromedaries, are lacking. Recent work has provided the first complete genome sequence of the dromedary (Wu *et al.* 2014), and additional genomes from its congeners have also been made available (Jirimutu *et al.* 2012; Burger & Palmieri 2014; Wu *et al.* 2014). These studies have identified candidate loci responsible for various adaptations to desert conditions, insulin resistance and camels' unique immune system. Although interspecific comparative genomics in camels have proven useful, little knowledge regarding the intraspecific variation, especially in dromedaries, exists. Large-scale analyses of genetic variation, or polymorphisms, within a species or population can uncover additional candidates for selection through dense genome scans of population divergence or hitchhiking (Ellegren 2008). For example, genome-wide analysis of single-nucleotide polymorphisms (SNPs) in cattle has identified loci linked to milk production traits (Pryce *et al.* 2010) and this knowledge has been implemented in breeding programmes designed to improve production traits through the process of genomic selection (reviewed by Hayes *et al.* 2009; Schefers & Weigel 2012). Furthermore, these genomic scans of polymorphism can inform assessments of demographic history, where population bottlenecks and small population sizes, often associated with mammalian megafauna, can obscure the ability to detect patterns of selection in genomes (Akey *et al.* 2004; Pool *et al.* 2010).

In this study, we sequenced and assembled a second genome for the dromedary. The individual's ('Waris') origin can be traced back to North Africa and the Canary Islands. Both regions are genetically distinct from populations in southern Arabia and further east, yet are indistinguishable from one another despite dromedaries from the Canary Islands having been isolated since the fifteenth century (Schulz *et al.* 2010). We quantitatively compared our dromedary genome and its demographic history with the existing reference and identified SNPs useful for future studies on the evolutionary and agricultural importance of this species. Finally, we comment on the data availability and transparency of bioinformatic methods for next-generation sequencing studies and present our methods and results consistent with current recommendations (Whitlock 2011).

Materials and methods

Sample collection, sequencing and assembly

Whole blood from a female dromedary named 'Waris' living at the First Austrian Camel Riding School in Eitental, Austria, was collected during a routine veterinary examination, and an aliquot was used for genomic DNA extraction with the MasterPure DNA Purification Kit (Epicenter, USA). The mother of Waris originated from the population on the Canary Islands, whereas the father was of North African origin. A 500-bp insert paired-end library and a 5-kb mate-pair library were prepared and sequenced using three lanes and one lane, respectively, on an Illumina HiSeq 2000 system (Illumina, USA). Preprocessing of the sequence reads included the removal of adapter sequences and removal of reads with >10% uncalled bases and/or >50% of bases with a Phred-scaled quality score <4. After preprocessing, all 100-bp (paired-end) and 50-bp (mate-pair) reads were retained as the set of 'raw' reads. We trimmed the 3′ end of all raw reads using a modified Mott algorithm in POPOOLATION v1.2.2 (Kofler *et al.* 2011) to a minimum quality score of 20 and a minimum length threshold of 50 bp and 30 bp for the paired-end and mate-pair reads, respectively.

We corrected the trimmed, paired-end reads for substitution sequencing errors using QUAKE v0.3.5 (Kelley *et al.* 2010). Salzberg *et al.* (2012) showed previously that the error correction of sequencing reads can greatly improve the *de novo* assembly of genomes, including genomes assembled using the program ABYSS (Simpson *et al.* 2009). QUAKE uses the distributions of infrequent and abundant *k*-mers to model the nucleotide error rates and subsequently corrects substitution errors. As input to QUAKE and again after error correction, we counted the frequency of 20-mers in the paired-end reads using DSK v1.6066 (Rizk *et al.* 2013). To estimate genome size, we divided the total number of error-free 20-mers by their peak coverage depth.

We assembled the genome using the trimmed and error-corrected paired-end reads with ABYSS v1.3.6. To determine the optimal *k*-mer length, we repeated the assembly using k = 40–88 in 8-bp increments. All scaffolding steps were performed using the trimmed mate-pair reads also in ABYSS, and only scaffolds longer than 500 bp were retained. We evaluated the completeness of each assembly using CEGMA v2.4 (Parra *et al.* 2007) with the '–mam' parameter for mammalian intron structure. CEGMA annotates highly conserved, core eukaryotic genes (CEGs) that should be present in the genome.

From the resulting assemblies, we selected two (the one with the fewest scaffolds, k = 48, and the one with the longest N50, k = 64) for further evaluation in REAPR v1.0 (Hunt *et al.* 2013). REAPR evaluates the accuracy of an assembly through the identification of small, local errors (single base substitutions and short insertions/deletions) and mis-assemblies (such as structural or scaffolding errors) using mapped, paired-end reads. One of the primary metrics calculated by REAPR is the fragment coverage distribution (FCD). This statistic is measured on a per-site basis and is the distribution of coverage depth for fragments (regions between the outermost ends of a set of properly paired reads) containing the base. The difference between the observed FCD and its theoretical distribution is the FCD error, and strings of bases with high FCD error indicate assembly mistakes (Hunt *et al.* 2013). The FCD error cut-off for calling a failed region was determined automatically in REAPR after randomly sampling 10^5 windows of 100 bp in length. The scaffolds are cleaved at these locations to produce a 'broken assembly' more useful for comparison. As recommended input into REAPR, we mapped the trimmed and error-corrected paired-end reads to each genome assembly using SMALT v0.7.0.1 (https://www.sanger.ac.uk/resources/software/smalt) with default parameters. To assess the effects of error-correcting reads prior to assembly, we repeated the assembly ($k = 48$ and $k = 64$), CEGMA and REAPR analyses as described above using the trimmed (uncorrected) paired-end reads.

We selected the assembly with the highest proportion of error-free bases, fewest FCD errors and the longest N50 in the broken assembly. We assessed the composition of the short (<500 bp) scaffolds omitted from the final assembly using a BLASTN v2.2.30 (http://ncbi.nlm.nih.gov/blast) search against the nucleotide database of GenBank with an e-value cut-off of 10^{-3}. The genome assembly is available in GenBank as Accession no. GCA_000803125.1.

Comparison with existing dromedary genome

We further assessed the quality of our genome assembly through comparison with the recently published dromedary reference (Wu *et al.* 2014) (GenBank Accession no. GCA_000767585.1). We downloaded the raw reads for the three short-insert libraries (170-, 500- and 800-bp inserts) from the reference assembly. As described in Wu *et al.* (2014), we removed reads with >5% uncalled bases, with >40 bases of Phred-scaled quality ≤20, with adapter contamination (match length ≥10 bp, mismatch ≤3 bp), with duplicated forward and reverse pairs and with overlapping forward and reverse pairs (excluding the 170-bp insert library, overlap ≥10 bp, mismatch ≤10% bp). We then error-corrected 17-mers that only occurred once (Wu *et al.* 2014) and repeated the REAPR pipeline separately for each library as described above.

In addition, we performed a separate whole-genome alignment of both complete dromedary genome assemblies using MUGSY v1.2.3 (Angiuoli & Salzberg 2011) with a maximum distance of 500 bp for chaining anchors into locally collinear blocks. The final alignment blocks were filtered using MAFFILTER v1.1.0 (Dutheil *et al.* 2014) with the following criteria: using a sliding window of 10 bp, we excluded the window from the alignment if more than five gaps (including 'N') were present and subsequently split the block. We retained alignment blocks with a minimum length of 500 bp.

Genome annotation

We employed a two-pass, iterative procedure using the MAKER v2.31.6 pipeline (Cantarel *et al.* 2008; Holt & Yandell 2011) to manage and evaluate the different evidences for gene annotation. For the first pass, we predicted genes using SNAP (Korf 2004) with hidden-Markov models developed from the CEGs identified from CEGMA and an *ab initio* prediction of genes from GENEMARK-ES (Lomsadze *et al.* 2005). This first pass also included alignments from existing dromedary ESTs (Al-Swailem *et al.* 2010) and protein-based homology from a concatenated set of Bactrian camel (Accession no. GCF_000311805.1), alpaca (Accession no. GCF_000164845.1) and cow (Accession no. GCF_000003055.4) protein sequences. For the second pass, we predicted genes using both SNAP and AUGUSTUS v2.5.5 (Stanke *et al.* 2006), both trained with a hidden-Markov model developed from the predictions of the first MAKER pass. The second pass also included the EST- and protein-based evidence as described in the first pass. All runs of MAKER included the masking of repetitive regions using REPEATMASKER v4.0.3 (Smit *et al.* 1996–2010) against the REPBASE v19.07 (Jurka *et al.* 2005) library. For each gene prediction, we selected the evidence with an annotation edit distance (AED) < 0.75.

Using the longest isoform for each protein sequence, we functionally annotated each gene using a combination of BLASTP v2.2.30 (http://ncbi.nlm.nih.gov/blast) and INTERPROSCAN 5.7.48 (Jones *et al.* 2014). BLAST searches were performed against metazoan protein sequences from the 'nr' database with an e-value cut-off of 10^{-3}, and only the top 20 hits were retained. We used INTERPROSCAN to assign protein domains and motif to sequences through comparison against a variety of databases (i.e. TIGRFAM, PRODOM, SMART, HAMAP, PROSITEPATTERNS, SUPERFAMILY, PRINTS, PANTHER, GENE3D, PIRSF, PFAMA, PROSITEPROFILES, COILS). Annotations were stored as Gene Ontology (GO) terms for each sequence. Next, we used the protein sequences to identify single-copy orthologs shared with the *C. ferus* (GCA_000311805.2) and with the *Bos taurus* (Accession no. GCF_000003055.5) genomes using ORTHOMCL (Li *et al.* 2003). We used a minimum identity of 30% and an e-value cut-off of 10^{-5} to call orthologs.

In combination with the homology-based repeat annotation described above, we also characterized *de novo* repetitive elements from the sequencing reads and genome assembly using separate approaches. To identify repeats directly from the trimmed and error-corrected paired-end reads, we used the method implemented in REPARK v1.2.1 (Koch *et al.* 2014). This method works by generating a *de novo* assembly of the abundant *k*-mers (k = 31) in the reads. REPARK determined the threshold for defining abundant 31-mers by fitting a linear function to the slope of the descending segment of the Poisson-like unique *k*-mer fraction (Fig. S1, Supporting information). The abundant 31-mers were defined as those occurring at frequency greater than twice the *x*-intercept of the linear function. The *x*-intercept of our linear function was 49, and therefore, abundant 31-mers were defined as those occurring more than 98 times in the sequencing reads. The abundant 31-mers were assembled with VELVET v2.0 (Zerbino & Birney 2008). We calculated statistics for the contigs using QUAST v2.3 (Gurevich *et al.* 2013). We identified and classified repeat families for both assemblies (the repetitive 31-mers and the genome assembly) using a combination of RECON v1.08 (Bao & Eddy 2002) and REPEATSCOUT v1.0.5 (Wootton & Federhen 1993; Benson 1999). Final repeat libraries for each assembly were subsequently built using REPEATMODELER v2.1 (Smit & Hubley 2008–2010).

The noncoding RNA genes were predicted with structure-based homology search by INFERNAL v1.1.1 (Nawrocki *et al.* 2009) against the RFAM database (Release 12.0) (Griffiths-Jones *et al.* 2003). We used a 'gathering' cut-off score of 85% for the covariance models and a confidence threshold (e-value) of 10^{-9}. We annotated CpG islands using the 'cpgplot' tool in EMBOSS v6.5.7 (Rice *et al.* 2000) with the repeat-masked genome employing a window length of 100 bp, a minimum island length of 200 bp, minimum GC content of 0.5 and a minimum average observed ratio of C+G to CpG of 0.6.

Variant identification and demographic analysis

We aligned the trimmed and error-corrected paired-end reads back to the final genome assembly using BWA v0.6.2 (Li & Durbin 2009). From the alignment, we removed duplicated reads and filtered all alignments to contain only unambiguously mapped and properly paired reads using SAMTOOLS v1.1 (Li *et al.* 2009). We identified variants (SNPs and insertion/deletion polymorphisms) using a combination of SAMTOOLS and PLATYPUS (Rimmer *et al.* 2014). Both of these variant callers have been shown to produce reliable results for single-sample SNP calling and do not require preprocessing steps that realign reads around indels and recalibrate base quality scores (Liu *et al.* 2012; Baes *et al.* 2014). As recommended by Baes *et al.* (2014), we included the consensus set of variants identified by both methods. We further excluded variants with a Phred-scaled quality score <20, that were within five base pairs of another variant, and whose depth of coverage was less than 1/3 or more than twice the mean genome coverage of the alignment. The quality of the final set of variants was assessed using the ratio of transitions (pyrimidine ↔ pyrimidine or purine ↔ purine) to transversions (purine ↔ pyrimidine) in VCFTOOLS v0.1.12b (Danecek *et al.* 2011). This ratio, called the ti/tv ratio, is known to be ~2.1 in human genomes and is often used to evaluate variant prediction quality (DePristo *et al.* 2011; Liu *et al.* 2012; Baes *et al.* 2014). The SNP density within the genome and divergent sites from alignment with the reference genome were estimated using nonoverlapping 1000-bp windows and then separately for the annotated regions (*i.e.* exons, introns, CpG islands, repetitive regions) in VCFTOOLS.

We examined the historical changes in effective population size (N_e) of the dromedary genome using the pairwise sequentially Markovian coalescent model (PSMC v0.6.4) (Li & Durbin 2011). PSMC infers N_e at a given time in the past from a single diploid individual using the rates of coalescence events across the genome. Because PSMC is highly dependent on the density of polymorphic sites, we performed two different runs of PSMC: (i) using only the sites with a mean mapping quality ≥20 and coverage between one-third and twice the mean genome coverage (lenient conditions) and (ii) a consensus genome sequence generated from our filtered set of variants described above and with repetitive regions masked (strict conditions). Both analyses in PSMC were performed for 25 iterations using -p and -t parameters chosen manually to infer ~10 recombination events in the interval (Li & Durbin 2011) and an initial theta/rho ratio ($-r$) of 5. The variance was assessed using 100 bootstrap replicates, and final estimates of N_e and time were scaled with a mutation rate of 2.5×10^{-8} and a generation time of five years.

Results and discussion

Sequencing and assembly comparisons

We sequenced the genome of a female dromedary of North African ancestry, 'Waris', using only one short-insert (500 bp) and one long-insert (5 kb) library. Prior to error-correcting reads, these shotgun libraries generated 66.4× coverage of the genome. A summary of the sequencing reads and estimated genome coverage can be found in Table 1. We counted the frequency of unique 20-mers in the trimmed paired-end reads and, using 20-mers with a frequency of three or less, determined the rate of base substitution error to be 2.7% (Table S1, Supporting information). This error rate is higher than that commonly reported for the Illumina HiSeq 2000 system (0.1%–1%) (Glenn 2011; Minoche *et al.* 2011) and may be the result of reduced sequencing performance and/or

Table 1 Read statistics after quality and length trimming

Library	# Reads with partner	# Reads without partner	Mean length (SD)	Total number of bases	Sequence coverage
500-bp PE	579 823 726	5 045 754	98.2 (6.4)	114 374 878 323	55.7×
500-bp PE-corrected	562 416 289	22 102 005	98.1 (7.0)	112 536 342 122	54.8×
5-kb MP	224 408 840	2 834 348	48.6 (1.8)	21 970 012 359	10.7×
Total (Corrected+MP)	786 825 129	24 936 353	—	134 506 354 481	65.5×

PE, paired-end library; MP, mate-pair library.

the presence of low-abundance, contaminating sequence (e.g. humans). For instance, the extracted DNA was from whole blood, which may contain a wide variety of microorganisms whose DNA abundance is rare relative to the host. As suggested by Salzberg *et al.* (2012), we corrected reads for these errors (Fig. S2, Supporting information) and found only a 1.6% reduction in the total number of bases used for assembly (~1× coverage reduction in the final assembly) (Table 1). A majority of corrections were made to bases with a Phred-scaled quality score <10 and were consistent between forward, reverse and unpaired reads (Fig. S3, Supporting information). Using the counts of 20-mers with a frequency >3 and a peak coverage of 35x, we estimated the genome size to be 2.25 gigabases (Gb). This estimate is similar to that reported previously for the dromedary (2.27 Gb) using the frequency of 17-mers (Wu *et al.* 2014) but less than that reported using flow cytometry (2.56 Gb; Krishan *et al.* 2005).

We compared different *k*-mer sizes for assembly of the trimmed and error-corrected paired-end reads (Fig. S4, Supporting information) and found that $k = 48$ produced the fewest scaffolds (24 058) and most CEGs (99.1%), whereas $k = 64$ produced the longest N50 (1 482 444 bp) and longest scaffold (9 719 801 bp). A quantitative comparison of these two assemblies both before and after error correction revealed that the use of error-corrected reads produced assemblies with more error-free bases and fewer gaps, FCD errors and collapsed repeats (Table S2, Supporting information). Error-correcting reads also generated broken assemblies with longer N50 values (Table S2, Supporting information). We selected the assembly using corrected reads and $k = 64$, which outperformed the other assemblies in a variety of metrics given in Table S2 (Supporting information) (e.g. most error-free bases, fewest FCD errors, fewest gaps, longest N50 in the broken assembly). The final assembly was 2.06 Gb and contained 35 752 scaffolds (≥500 bp) with a GC content of 41.3% (Table 2). We omitted ~4.1 million small scaffolds (<500 bp) from the assembly, a majority (66.6%) of which either had no databases matches or were excluded from searches by the default low-complexity filter in BLAST. Of the remaining small scaffolds with a database match, ~1.1 million (80.2%) were *C. dromedarius* microsatellite sequences and the rest were distributed among other species, especially *Vicugna pacos*, *Sus scrofa* and *Homo sapiens* (Fig. S5, Supporting information). The N50 of the assembly was 1.48 megabases (Mb), and 95% of the assembly was contained in the longest 2379 scaffolds (Fig. 1). We annotated 452 (98.7%) CEGs, indicative of the completeness of the assembly.

Table 2 Summary of the dromedary genome assembly presented in this study compared with the current reference

	k = 64-C African dromedary	Reference* Arabian dromedary
# Scaffolds	35 752	32 572
Mean length (bp)	57 481.1	61 526.7
Total length (bp)	2 055 063 633	2 004 047 047
Longest (bp)	9 719 801	23 736 781
GC content	41.3%	41.2%
Repeat content	33.7%	28.4%
N50 (count)	1 482 444 (393)	4 188 677 (132)
N60 (count)	1 108 832 (553)	2 993 967 (190)
N70 (count)	842 144 (764)	2 137 136 (268)
N80 (count)	558 658 (1063)	1 311 427 (389)
N90 (count)	260 185 (1592)	689 795 (594)
Number of gaps	150 386	72 775
Total gap length	53 439 631	22 596 073
CEGs†	98.7%	98.5%

*Accession no. GCA_000767585.1; Wu *et al.* (2014).
†Proportion of 458 core eukaryotic genes (CEGs) identified using CEGMA.

Many assembly characteristics (e.g. number of scaffolds, mean scaffold length, GC content, repeat content, CEGs identified) were markedly similar to the current dromedary reference genome, suggesting that the *C. dromedarius* genome sequences are relatively robust to the assembly method used. Our assembly did have a shorter scaffold distribution than the current reference (N50 = 1.48 Mb compared with 4.2 Mb, respectively) and contained twice as many gaps (150 386 compared with 72 775, respectively) (Table 2). Because N50 and other scaffold length metrics are not necessarily indicative of assembly quality (Bradnam *et al.* 2013; Hunt

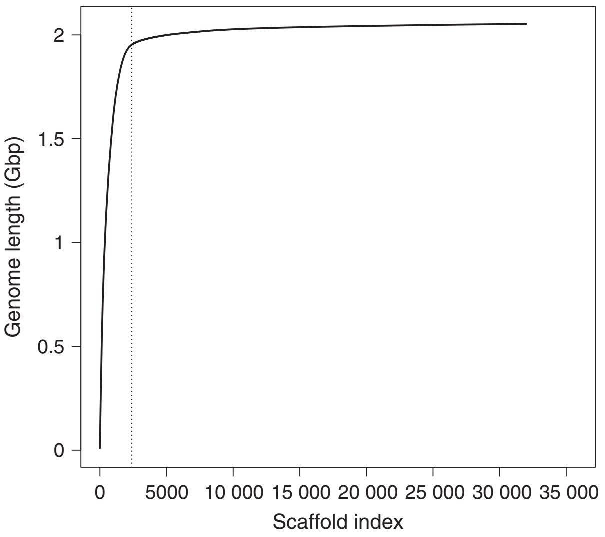


Fig. 1 Cumulative length of the African *Camelus dromedarius* assembly. Scaffolds are sorted from longest to smallest along the horizontal axis. The vertical dotted line indicates the number of scaffolds containing 95% of the total assembly.

et al. 2013), we quantitatively compared our assembly with the existing reference using trimmed and error-corrected paired-end reads mapped to the genome sequence (see Table S3, Supporting information for read Accession nos). Our genome consistently had a larger proportion of error-free bases (91.8%), fewer FCD errors (37 015) and fewer reads in the wrong orientation (113 677), whereas the reference assembly often contained fewer collapsed repeats (Table 3). Furthermore, the cut-off for defining FCD errors in our assembly was more stringent than in the reference, and when comparing the same cut-off, fewer windows were called as errors (Fig. 2). These results support that traditional assembly statistics (e.g. N50, mean length, number of scaffolds) do not necessarily indicate the quality and suggest that more robust quantitative comparisons should be performed. For example, the method employed by Wu *et al.* (2014) to assemble the reference genome has been shown to artificially inflate scaffold lengths at the expense of increasing assembly errors (Salzberg *et al.* 2012; Bradnam *et al.* 2013).

An alignment of the two genomes produced 291 611 blocks with a total alignment length of 1.91 Gb. The mean block length was 6549.4 bp (SD 7006.0 bp) (Fig. S6, Supporting information). This result further supports the high degree of similarity between the genome sequences despite different assembly strategies.

Genome annotation

We utilized a combination of *ab initio* and evidence-based homology to identify and annotate protein-coding elements in the genome. Not accounting for multiple isoforms, we predicted a total of 21 167 genes containing either protein- or EST-based evidence (Fig. S7, Supporting information); a number similar to that reported for the dromedary reference genome (20 714; Wu *et al.* 2014). Nearly, all genes (98.7%) returned a significant match to known metazoan protein sequences, often with high similarity (Fig. S8A, Supporting information). A majority of the top hits for each gene matched other camelid sequences such as *C. ferus* (57.9%) and *Vicugna*

Table 3 Frequency of different assembly errors compared with the reference genome for short-insert reads (separated by insert size)

	k = 64-C African dromedary	Reference* Arabian dromedary		
Insert size	500 bp	170 bp	500 bp	800 bp
Error-free bases	91.8%	83.4%	74.9%	68.6%
FCD† errors	37 015	9 641 002	203 806	195 429
Collapsed repeats	10 233	86 488	8694	4659
Wrong read orientation	113 677	95 230	215 821	210 951

*Accession no. GCA_000767585.1; Wu *et al.* (2014).
†Fragment coverage distribution.

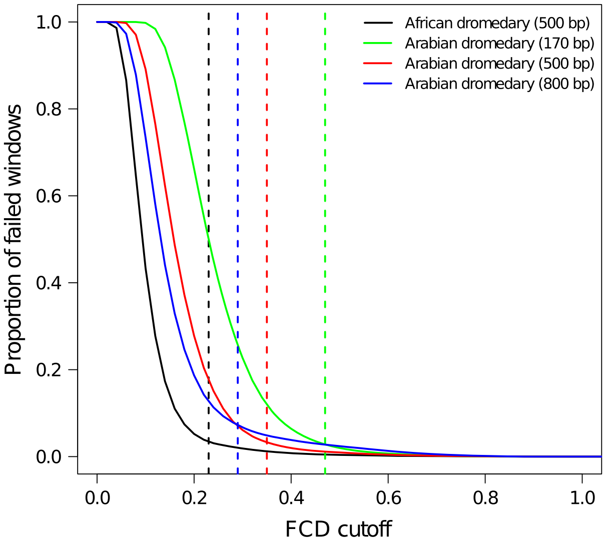


Fig. 2 Calculation of the fragment coverage distribution (FCD) error cut-off. For each potential FCD cut-off, each solid line represents the proportion of 100-bp windows that would fail and subsequently be labelled as an assembly error. The vertical dashed lines are the cut-off scores determined in REAPR using the value where the normalized (between −1 and 1) first and second derivatives are ≥0.05. See Hunt *et al.* (2013) for a complete description of the method. The colours correspond with the different read alignments separated by genome and insert size.

pacos (14.1%) (Fig. S8B, Supporting information). We added functional annotations to 17 779 (84.0%) sequences using INTERPROSCAN. A total of 32 965 GO terms were also mapped to 13 198 sequences (Fig. S9, Supporting information). Both the number of single-copy orthologs and their mean amino acid identity were higher when compared with the *C. ferus* genome (12 170 and 95.1%, respectively) than when compared with the *B. taurus* genome (11 625 and 86.3%, respectively) (Fig. S10, Supporting information). A comparison with the dromedary reference genome was not possible because, at the time of writing, annotation data remained unavailable. Because the annotation pipeline we used was designed to promote future reannotation as more data become available, the accuracy in gene predictions can easily be increased over time.

We estimated 33.7% of our genome to be composed of repetitive elements using a combination of homology-based and *de novo* approaches (Tables S4 and S5 and Fig. S11, Supporting information). The homology-based search identified 31.7% of the genome as repetitive, whereas the *de novo* methods based upon the sequencing reads or the assembly predicted less (13.23% and 24.39%, respectively). Only ~2% of the combined set of repetitive elements were specific to the *de novo* approaches which included primarily LINE1 retrotransposons and unclassified repeats (Table S5, Supporting information). Overall, LINE elements accounted for 19.3% of the genome (Fig. S11, Supporting information). We found a total of 3691 noncoding RNA loci (Table S6, Supporting information), including 1369 micro RNAs, 966 small nuclear RNAs and 524 small nucleolar RNAs. We classified 57 708 putative CpG islands that had a mean length of 326.3 (SD 154.1) bases.

Variant identification and demographic analysis

We mapped 94.1% of the trimmed and error-corrected paired-end reads to our genome assembly. After quality control and filtering, 75.8% of the reads were retained resulting in a mean alignment coverage of 40.8x. We identified a set of ~1.4 millions SNPs and 162 538 insertion/deletion polymorphisms that overlapped between the two SNP-calling algorithms and passed our filtering criteria. The ti/tv ratio for our final set of SNPs was 2.31, consistent with the ratio reported in dairy cattle using the same algorithms and characteristic of a low rate of false-positive SNPs (DePristo *et al.* 2011; Baes *et al.* 2014).

Across the genome, mean SNP density (heterozygosity) was 0.71×10^{-3} (SD 1.4×10^{-3}), slightly less than reported for the Arabian dromedary (0.74×10^{-3}; Wu *et al.* 2014). This reduction may be the result of either technical differences in SNP calling (e.g. the method or filtering criteria used) or the consequence of demographic events (e.g. smaller effective population size, increased inbreeding) in the North African/Canary Island population. We suspect the former, considering that for microsatellite data, the Arabian dromedary has a higher F_{IS} and lower levels of both observed heterozygosity and allelic richness than dromedaries from North Africa and the Canary Islands (Schulz *et al.* 2010). Nonetheless, SNP density in dromedaries appears to be substantially less than that reported for domestic Bactrian and wild camels (1.0–1.29 × 10^{-3}; Jirimutu *et al.* 2012; Burger & Palmieri 2014; Wu *et al.* 2014). Within the dromedary genome, SNP density was highest in CpG islands (0.88×10^{-3}). This is consistent with the hypermutability of CpG residues (Coulondre *et al.* 1978; Ehrlich & Wang 1981; Sved & Bird 1990) and the positive relationship between mutation rate and CpG content (Walser & Furano 2010). SNP density was lowest in exons (0.47×10^{-3}) and at intermediate levels in both introns (0.57×10^{-3}) and repetitive elements (0.64×10^{-3}) (Fig. 3). Because we omitted SNPs with excessively high coverage, SNP density in repetitive regions may be underestimated.

An alignment with the reference genome produced more than 1.7 million divergent sites, of which 631 468 (36.1%) were biallelic SNPs and the remaining sites were either insertion–deletion polymorphisms or uncalled bases. Nearly, all of these biallelic SNPs (99.4%) overlapped with the SNPs identified within our genome assembly. The relative density of divergent sites across different elements of the genome was similar to the

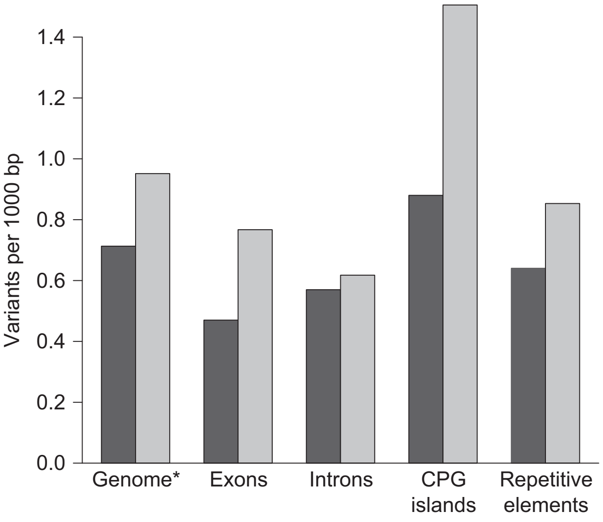


Fig. 3 Density of SNPs within the African dromedary genome assembly (dark grey bars) and density of divergent sites (light grey bars) from the alignment with the reference genome (Accession no. GCA_000767585.1). * Genome-wide density is based upon 1000-bp nonoverlapping windows.

density of SNPs (Fig. 3), with the exception of introns, which contained fewer divergent sites (0.62×10^{-3}) than exons (0.77×10^{-3}). Because introns are expected to contain more variation than exons, this result may be the product of increased alignment ambiguity and subsequent filtering of the more variable regions. The density of divergent sites was also markedly higher within CpG islands (1.51×10^{-3}) than in all other genomic regions (Fig. 3).

We examined the historical demography using the PSMC model and found consistent histories with little variance among both lenient and strict conditions (Fig. 4). Both conditions had a maximum N_e of ~20 000 approximately 350 thousand years before present (kybp) with a substantial bottleneck suffered thereafter. This bottleneck reduced N_e by nearly 70% during the ~ 300–100 kybp interval leading up to the last glacial period (LGP). The N_e declined gradually during the LGP between 100 and 20 kybp. At this time, at the end of the last glacial maxima (LGM; ~20 kybp), the lenient and strict conditions indicated either a small increase or constant N_e, respectively, until a second, more recent, bottleneck further reduced N_e to <1000 individuals beginning 10 kybp. The number of coalescent events occurring more recently than ~1 kybp is inadequate to accurately infer demographic history from this period. This pattern of climate-driven demographic changes has been observed in a variety of mammalian megafauna (Lorenzen *et al.* 2011; Orlando *et al.* 2013; Wu *et al.* 2014), although anthropogenic effects may have played a critical role in the most recent population reduction. More extensive surveys of camel remains in the archaeological record would be required to disentangle the roles of climate change and humans in driving the decline in dromedary population size. Unfortunately, Wu *et al.* (2014) did not report the generation time or mutation rate used to scale the demographic history of the Arabian dromedary, thus preventing a more thorough comparison with our result.

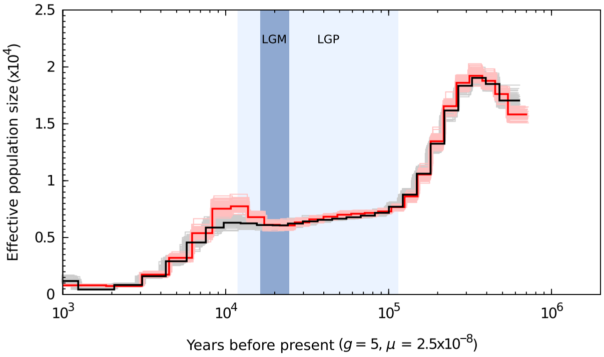


Fig. 4 Historical effective population size of the African dromedary inferred with the filtered, repeat-masked set of variants (black line; strict conditions) and with default parameters (red line; lenient conditions) in PSMC (Li & Durbin 2011). The lighter-coloured lines of the same colour represent the 100 bootstrap replicates. The result is scaled using a generation time (g) of five years and a per-base mutation rate (μ) of 2.5×10^{-8}. The light-blue and blue-shaded regions indicate the last glacial period (LGP) and last glacial maximum (LGM), respectively.

Conclusions

Here, we reported a second dromedary genome sequence that provides additional genetic resources from a geographically distinct region. Our results demonstrated that draft genome assemblies constructed using only one short- and one long-insert sequencing libraries can be comparable to those sequenced using more than two library sizes (e.g. 6 in this comparison). This suggests that rather than sequencing numerous libraries of various sizes resources are better spent on physically mapping the genome or on different technologies. For example, methods such as optical mapping (Chamala *et al.* 2013; Shearer *et al.* 2014) or long-read sequencing (Huddleston *et al.* 2014; Laszlo *et al.* 2014) have proven useful to improve the assembly of complex regions or otherwise finish draft genome sequences.

Many comparisons of our genome annotations (e.g. SNPs, coding sequences, noncoding RNAs) with the current dromedary reference genome were not possible due to the unavailability of these data. Therefore, in congruence with current recommendations for data sharing in ecology and evolution (Whitlock 2011), we have archived all data for this study in various locations (see Data accessibility section below) thus providing extensive resources to the camel-research community. In addition to the data, we make example bioinformatics code available to promote open, reproducible research and external evaluation as advocated by others (Mesirov 2010; Stodden *et al.* 2010; Peng 2011; Groves & Godlee 2012). The availability of genomic resources for dromedaries will facilitate future evolutionary studies of camels and the application of marker-assisted breeding selection to improve the yield and performance of camel-derived products. Because camelids, notably dromedaries, are especially adapted to harsh, arid environments, understanding how the process of natural and artificial selection that has shaped their unique traits has implications in both evolutionary biology and agriculture.

Acknowledgements

We are extremely grateful to Gerda Gassner from the First Austrian Camel Riding School for allowing a sample from Waris to be used for genome sequencing. We wish to acknowledge CSC – IT Centre for Science, Finland for generous computational resources. We thank Huiguang Wu, Inner Mongolia Cultural University (PRC) for valuable discussions and Christian

Schlötterer, Institute of Population Genetics, Vetmeduni Vienna (AUT) for general support. JC and the computational work in this research were supported by the grant 251170 from Academy of Finland to COIN centre of excellence. This work, RRF and EM were supported by the Austrian Science Foundation (FWF): P24706-B25 to PAB, recipient of an APART fellowship (11506) of the Austrian Academy of Sciences.

References

Agrawal RP, Jain S, Shah S, *et al.* (2011) Effect of camel milk on glycemic control and insulin requirement in patients with type 1 diabetes: 2-years randomized controlled trial. *European Journal of Clinical Nutrition*, **65**, 1048–1052.

Akey JM, Eberle MA, Rieder MJ, *et al.* (2004) Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biology*, **2**, e286.

Al-Swailem AM, Shehata MM, Abu-Duhier FM, *et al.* (2010) Sequencing, analysis, and annotation of expressed sequence tags for *Camelus dromedarius*. *PLoS ONE*, **5**, e10720.

Angiuoli SV, Salzberg SL (2011) Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics*, **27**, 334–342.

Baes CF, Dolezal MA, Koltes JE, *et al.* (2014) Evaluation of variant identification methods for whole genome sequencing data in dairy cattle. *BMC Genomics*, **15**, 948.

Bao Z, Eddy SR (2002) Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Research*, **12**, 1269–1276.

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, **27**, 573–580.

Bradnam KR, Fass JN, Alexandrov A, *et al.* (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, **2**, 10.

Bulliet R (1990) *The Camel and the Wheel*. Columbia University Press, New York.

Burger PA, Palmieri N (2014) Estimating the population mutation rate from a *de novo* assembled Bactrian camel genome and cross-species comparison with Dromedary ESTs. *Journal of Heredity*, **105**, 839–846.

Cantarel BL, Korf I, Robb SM, *et al.* (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, **18**, 188–196.

Chamala S, Chanderbali AS, Der JP, *et al.* (2013) Assembly and validation of the genome of the nonmodel basal angiosperm *Amborella*. *Science*, **342**, 1516–1517.

Coulondre C, Miller JH, Farabaugh PJ, Gilbert W (1978) Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature*, **274**, 775–780.

Danecek P, Auton A, Abecasis G, *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

DePristo MA, Banks E, Poplin R, *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–498.

Dutheil JY, Gaillard S, Stukenbrock EH (2014) MafFilter: a highly flexible and extensible multiple genome alignment files processor. *BMC Genomics*, **15**, 53.

Ehrlich M, Wang RYH (1981) 5-methylcytosine in eukaryotic DNA. *Science*, **212**, 1350–1357.

Ellegren H (2008) Comparative genomics and the study of evolution by natural selection. *Molecular Ecology*, **17**, 4586–4596.

Glenn TC (2011) Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, **11**, 759–769.

Griffiths-Jones S, Bateman A, Marshall M, *et al.* (2003) Rfam: an RNA family database. *Nucleic Acids Research*, **31**, 439–441.

Grigson C (2012) Camels, copper and donkeys in the early Iron Age of the southern Levant: Timna revisited. *Levant*, **44**, 82–100.

Groves T, Godlee F (2012) Open science and reproducible research. *BMJ*, **344**, e4383.

Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, **29**, 1072–1075.

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Invited review: genomic selection in dairy cattle: progress and challenges. *Journal of Dairy Science*, **92**, 433–443.

Holt C, Yandell M (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, **12**, 491.

Huddleston J, Ranade S, Malig M, *et al.* (2014) Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Research*, **24**, 688–696.

Hunt M, Kikuchi T, Sanders M, *et al.* (2013) REAPR: a universal tool for genome assembly evaluation. *Genome Biology*, **14**, R47.

Jirimutu, Wang Z, Ding G, *et al.* (2012) Genome sequences of wild and domestic Bactrian camels. *Nature Communications*, **3**, 1202.

Jones P, Binns D, Chang HY, *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.

Jurka J, Kapitonov VV, Pavlicek A, *et al.* (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, **110**, 462–467.

Kelley DR, Schatz MC, Salzberg SL (2010) Quake: quality-aware detection and correction of sequencing errors. *Genome Biology*, **11**, R116.

Koch P, Platzer M, Downie BR (2014) RepARK—de novo creation of repeat libraries from whole-genome NGS reads. *Nucleic Acids Research*, **42**, e80.

Kofler R, Orozco-terWengel P, De Maio N, *et al.* (2011) PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS ONE*, **6**, e15925.

Köhler-Rollefson IU (1991) *Camelus dromedarius*. *Mammalian Species*, **375**, 1–8.

Korf I (2004) Gene finding in novel genomes. *BMC Bioinformatics*, **5**, 59.

Krishan A, Dandekar P, Nathan N, *et al.* (2005) DNA index, genome size, and electronic nuclear volume of vertebrates from the Miami Metro Zoo. *Cytometry Part A*, **65A**, 26–34.

Laszlo AH, Derrington IM, Ross BC, *et al.* (2014) Decoding long nanopore sequencing reads of natural DNA. *Nature Biotechnology*, **32**, 829–833.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

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

Li L, Stoeckert CJ Jr, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, **13**, 2178–2189.

Li H, Handsaker B, Wysoker A, *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Liu Q, Guo Y, Li J, *et al.* (2012) Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. *BMC Genomics*, **13**(Suppl. 8), S8.

Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Research*, **33**, 6494–6506.

Lorenzen ED, Nogúes-Bravo D, Orlando L, *et al.* (2011) Species-specific responses of Late Quaternary megafauna to climate and humans. *Nature*, **479**, 359–364.

Mesirov JP (2010) Accessible reproducible research. *Science*, **327**, 415–416.

Minoche AE, Dohm JC, Himmelbauer H (2011) Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biology*, **12**, R112.

Muyldermans S, Baral TN, Retamozzo VC, *et al.* (2009) Camelid immunoglobulins and nanobody technology. *Veterinary Immunology and Immunopathology*, **128**, 178–183.

Nawrocki EP, Kolbe DL, Eddy SR (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**, 1335–1337.

Orlando L, Ginolhac A, Zhang G, *et al.* (2013) Recalibrating *Equus* evolution using the genome sequence of an early Middle Pleistocene horse. *Nature*, **499**, 74–78.

Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, **23**, 1061–1067.

Peng RD (2011) Reproducible research in computational science. *Science*, **334**, 1226–1227.

Pool JE, Hellmann I, Jensen JD, Nielsen R (2010) Population genetic inference from genomic sequence variation. *Genome Research*, **20**, 291–300.

Pryce JE, Bolormaa S, Chamberlain AJ, *et al.* (2010) A validated genome-wide association study in 2 dairy cattle breeds for milk production and fertility traits using variable length haplotypes. *Journal of Dairy Science*, **93**, 3331–3345.

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics*, **16**, 276–277.

Rimmer A, Phan H, Mathieson I, *et al.* (2014) Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics*, **46**, 912–918.

Rizk G, Lavenier D, Chikhi R (2013) DSK: k-mer counting with very low memory usage. *Bioinformatics*, **29**, 652–653.

Salzberg SL, Phillippy AM, Zimin A, *et al.* (2012) GAGE: a critical evaluation of genome assemblies and assembly algorithms. *Genome Research*, **22**, 557–567.

Schefers JM, Weigel KA (2012) Genomic selection in dairy cattle: integration of DNA testing into breeding programs. *Animal Frontiers*, **2**, 4–9.

Schulz U, Tupac-Yupanqui I, Martínez A, *et al.* (2010) The Canarian camel: a traditional dromedary population. *Diversity*, **2**, 561–571.

Shearer LA, Anderson LK, de Jong H, *et al.* (2014) Fluorescence *in situ* hybridization and optical mapping to correct scaffold arrangement in the tomato genome. *G3-Genes Genomes Genetics*, **4**, 1395–1405.

Simpson JT, Wong K, Jackman SD, *et al.* (2009) ABySS: a parallel assembler for short read sequence data. *Genome Research*, **19**, 1117–1123.

Smit AFA, Hubley R (2008–2010) RepeatModeler Open-1.0. Available from http://www.repeatmasker.org.

Smit AFA, Hubley R, Green P (1996–2010) RepeatMasker Open-3.0. Available from http://www.repeatmasker.org.

Spencer PBS, Woolnough AP (2010) Assessment and genetic characterisation of Australian camels using microsatellite polymorphisms. *Livestock Science*, **129**, 241–245.

Stanke M, Keller O, Gunduz I, *et al.* (2006) AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research*, **34**, W435–W439.

Stodden V, Donoho D, Fomel S, *et al.* (2010) Reproducible research. *Computing in Science and Engineering*, **12**, 8–13.

Sved J, Bird A (1990) The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proceedings of the National Academy of Sciences of the United States of America*, **87**, 4692–4696.

Uerpmann H-P, Uerpmann M (2002) The appearance of the domestic camel in south-east Arabia. *The Journal of Oman Studies*, **12**, 235–260.

Walser JC, Furano AV (2010) The mutational spectrum of non-CpG DNA varies with CpG content. *Genome Research*, **20**, 875–882.

Whitlock MC (2011) Data archiving in ecology and evolution: best practices. *Trends in Ecology and Evolution*, **26**, 61–65.

Wootton JC, Federhen S (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Computers and Chemistry*, **17**, 149–163.

Wu H, Guang X, Al-Fageeh MB, *et al.* (2014) Camelid genomes reveal evolution and adaptation to desert environments. *Nature Communications*, **5**, 5188.

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, **18**, 821–829.

R.R.F. wrote the study and performed bioinformatic analyses. E.M. extracted DNA, performed bioinformatic analyses and revised the manuscript. J.C. provided the extensive computational resources necessary for the completion of the project and revised the manuscript. P.A.B. managed the project, carried out initial raw data analysis and revised the manuscript.

Data accessibility

Project Information: NCBI BioProject PRJNA269274; NCBI SRA SRP050586.

Sample Information: NCBI BioSample – SAMN03252735; NCBI SRA – SRS779886.

Raw Sequence Reads: NCBI SRA – SRX796513 (500-bp insert), SRX796571 (5-kb insert).

Trimmed and Error-corrected paired-end reads: NCBI SRA – SRX1013838.

Genome Assembly: GenBank Accession no. GCA_000803125.1.

Read alignments (.bam format): NCBI SRA – SRR1950615.

SNP data, protein and RNA annotations, and the genome alignment are available in Dryad under doi:10.5061/dryad.v28f9.

Example scripts and code: Online Supporting Information, Methods S1 and S2.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1 The count of unique 31-mers that are found *n* (multiplicity) times in the trimmed and error-corrected paired-end reads (blue line).

Fig. S2 The count (A) and cumulative proportion (B) of unique 20-mers that are found *n* (multiplicity) times in the raw, paired-end sequencing reads (red line) and the trimmed and error-corrected paired-end reads (blue line).

Fig. S3 Histogram of the base quality scores corrected in the forward (green line), reverse (blue line) and unpaired (yellow line) raw reads.

Fig. S4 Comparison of (A) the number of scaffolds, (B) N50 length, (C) Proportion of 248 core eukaryotic genes (CEGs) annotated, and (D) longest scaffold length for various *k*-mer sizes used to assemble the genome.

Fig. S5 Distribution of the species (outer circle) and sequence types (inner circle) for the top blast hit for each of the short scaffolds (<500 bp) omitted from the final assembly.

Fig. S6 Histogram of the lengths (in base-pairs) of alignment blocks between our dromedary genome assembly and the reference (Accession no. GCA_000767585.1).

Fig. S7 Cumulative number of genes ordered by increasing AED (annotation edit distance) scores.

Fig. S8 The distribution of (A) similarity scores for all the BLAST hits and (B) the species of the top hit for each annotated protein sequence.

Fig. S9 The number of Gene Ontology (GO) terms mapped to each protein sequence.

Fig. S10 Histogram of the amino acid identity of single-copy orthologs between the African dromedary assembly and the

Camelus ferus (solid line) and *Bos taurus* (dashed line) genome assemblies.

Fig. S11 The relative abundance of repeat classes in the dromedary genome assembly vs. the Kimura divergence from the consensus, using the combined set of annotated repetitive elements.

Table S1 Summary statistics of unique 20-mers in the trimmed sequencing reads.

Table S2 Summary of the *de novo* assemblies made using both uncorrected and error-corrected (-C) reads.

Table S3 The accession numbers of raw reads for the dromedary reference (Accession no. GCA_000767585.1) downloaded and used for comparison.

Table S4 Statistics of the *de novo* assembled contigs from the abundant 31-mers.

Table S5 Statistics of the repetitive elements identified from *de novo* identification in the sequencing reads and the genome assembly, in addition to the homology-based search and the combined results.

Table S6 Summary of the non-coding RNA annotations in the dromedary genome assembly.

Methods S1 Example commands used for different analyses in this study.

Methods S2 Configuration files for the first (A) and second (B) iterations of MAKER v2.31.6.

Supporting Information for Online Publication

Supplementary Tables

Table S1. Summary statistics of unique 20-mers in the trimmed sequencing reads. N = number of unique *k*-mers; $N_{>3}$ = number of 20-mers occurring more than 3 times; $N_{<3}$ = number of 20-mers occurring less than 3 times; Error rate = $N_{<3}$ / (20*N); Total count = count of all 20-mers.

N	**$N_{>3}$ (True *k*-mers)**	**$N_{<3}$ (Error *k*-mers)**	**Error Rate**	**Total Count**
4,022,537,264	1,869,966,353	2,152,570,911	0.027	78,706,145,111

Table S2. Summary of the *de novo* assemblies made using both uncorrected and error-corrected (-C) reads. The "Broken Assembly" is the assembly created using REAPR after splitting scaffolds when an error occurs over a gap, or when an error contains more than one gap.

Assembly	***k*=48**	***k*=48-C**	***k*=64**	***k*=64-C**[1]
# Scaffolds	24,645	24,058	36,126	35,752
Mean Length	81,938.5	83,900.2	56,881.5	57,481.1
Total Length	2,019,374,771	2,018,471,819	2,054,899,881	2,055,063,633
Longest	8,065,708	5,799,438	7,585,309	9,719,801
N50 (count)	1,280,066 (472)	1,289,612 (469)	1,481,317 (415)	1,482,444 (393)
N60 (count)	982,628 (652)	992,344 (649)	1,102,948 (577)	1,108,832 (553)
N70 (count)	740,637 (889)	740,619 (886)	827,870 (794)	842,144 (764)
N80 (count)	512,933 (1,214)	506,582 (1,214)	551,192 (1,096)	558,658 (1,063)
N90 (count)	257,539 (1,755)	261,637 (1,754)	254,921 (1,642)	260,185 (1,592)
Number of Gaps	159,154	152,126	167,153	150,386
Total Gap Length	65,421,832	64,720,402	54,478,522	53,439,631
Error Free Bases	89.6%	89.6%	91.0%	91.8%
FCD Errors	48,604	48,050	44,673	37,015
Collapsed Repeats	8,479	8,411	10,707	10,233
Wrong Read Orientation	178,184	175,162	118,250	113,677
CEGs[1]	99.1%	99.1%	98.5%	98.7%
Broken Assembly				
# Scaffolds	57,049	55,758	69,307	62,302
Mean Length	35,349.4	36,152.5	29,603.7	32,947.9
Total Length	2,016,645,998	2,015,790,090	2,051,743,927	2,052,720,440
Longest	838,048	715,267	1,293,716	1,086,943
N50 (count)	103,294 (5,956)	106,329 (5,786)	150,381 (3,983)	174,415 (3,471)
N60 (count)	83,435 (8,133)	85,164 (7,907)	119,057 (5,514)	138,871 (4,785)
N70 (count)	64,995 (10,863)	66,211 (10,585)	89,475 (7,494)	105,027 (6,491)
N80 (count)	46,750 (14,500)	47,705 (14,157)	61,826 (10,238)	72,012 (8,841)
N90 (count)	26,601 (20,101)	27,221 (19,638)	32,223 (14,702)	37,287 (12,690)
Number of Gaps	144,064	137,838	148,697	136,068
Total Gap Length	70,606,453	69,951,648	58,361,196	56,665,254

[1] proportion of 458 core eukaryotic genes (CEGs) identified using CEGMA.

Table S3. The accession numbers of raw reads for the dromedary reference (Accession no. GCA_000767585.1) downloaded and used for comparison.

Insert Size		
170 bp	**500 bp**	**800 bp**
SRR1555056	SRR1555055	SRR1555057
SRR1555064	SRR1555078	SRR1555058
SRR1555066	SRR1555084	SRR1555059
SRR1555068	SRR1555085	SRR1555061
SRR1555071	SRR1555091	SRR1555095
SRR1555073	SRR1555093	SRR1555067
SRR1555075		SRR1555069
SRR1555079		SRR1555087
SRR1555083		SRR1555089

Table S4: Statistics of the *de novo* assembled contigs from the abundant 31-mers.

Contigs	**Size**
≥31 bp	83,811
Total Length	10,037,583 bp
≥1000 bp	754
Total Length	1,308,401 bp
Longest Contig	16,414 bp
GC %	38.72
N50 (count)	124 (15,769)
N75 (count)	74 (42,480)

Table S5. Statistics of the repetitive elements identified from *de novo* identification in the sequencing reads and the genome assembly, in addition to the homology-based search and the combined results.

	***de novo*: reads**		***de novo*: assembly**		**Homology**		**Combined**	
	Count	**(%)**	**Count**	**(%)**	**Count**	**(%)**	**Count**	**(%)**
SINEs	217,995	1.05	227,886	1.58	470,778	3.42	473,387	3.43
ALUs	0	0	0	0	7	0	7	0
MIRs	124,243	0.62	227,886	1.58	463,899	3.38	463,927	3.38
LINEs:	512,055	8.89	792,308	14.72	812,338	18.01	1,009,426	19.28
LINE1	506,055	8.87	648,810	13.42	446,105	13.30	642,633	14.57
LINE2	6,000	0.01	135,151	1.23	311,570	4.10	312,130	4.10
L3/CR1	0	0	6,809	0.04	40,821	0.44	40,821	0.44
LTR Elements:	116,420	1.24	268,540	4.33	294,297	5.21	324,636	5.43
ERVL	5,797	0.10	68,512	1.24	80,909	1.72	80,984	1.72
ERVL-MaLRs	62,307	0.77	123,436	1.67	137,977	2.35	138,020	2.35
ERV_class I	47,897	0.36	75,572	1.40	52,135	0.85	81,938	1.07
ERV_class II	419	0	529	0.01	153	0	571	0
DNA Elements:	24,454	0.20	224,646	2	339,366	3.43	341,448	3.44
hAT-Charlie	0	0	136,399	1.12	186,819	1.79	186,819	1.79
TcMar-Tigger	24,439	0.20	34,089	0.50	64,838	0.80	66,902	0.81
Unclassified	30,978	0.52	22,614	0.61	6,183	0.05	67,373	0.57
Small RNA	0	0	1,201	0.01	96,958	0.34	96,958	0.34
Satellites	0	0	0	0	264	0.00	264	0.00
Simple Repeats	535,838	1.09	458,937	0.95	502,727	1.02	504,900	1.03
Low Complexity	98,997	0.24	82,953	0.20	86,933	0.21	87,448	0.21
Total		**13.23**		**24.39**		**31.68**		**33.72**

Table S6. Summary of the non-coding RNA annotations in the dromedary genome assembly.

RNA class	**Loci**	**Models**
cis-reulatory RNA	24	18
lncRNA	21	21
miRNA	1,369	201
Ribozyme	1	1
rRNA	45	3
snoRNA	524	208
snRNA	966	11
tRNA	475	2
Other	266	178
Total	3,691	643

Figure S1. The count of unique 31-mers that are found *n* (multiplicity) times in the trimmed and error-corrected paired-end reads (blue line). The black and red dashed lines indicate the threshold for defining abundant 31-mers (*n* > 98) and the linear function fit to the descending region of the curve, respectively.

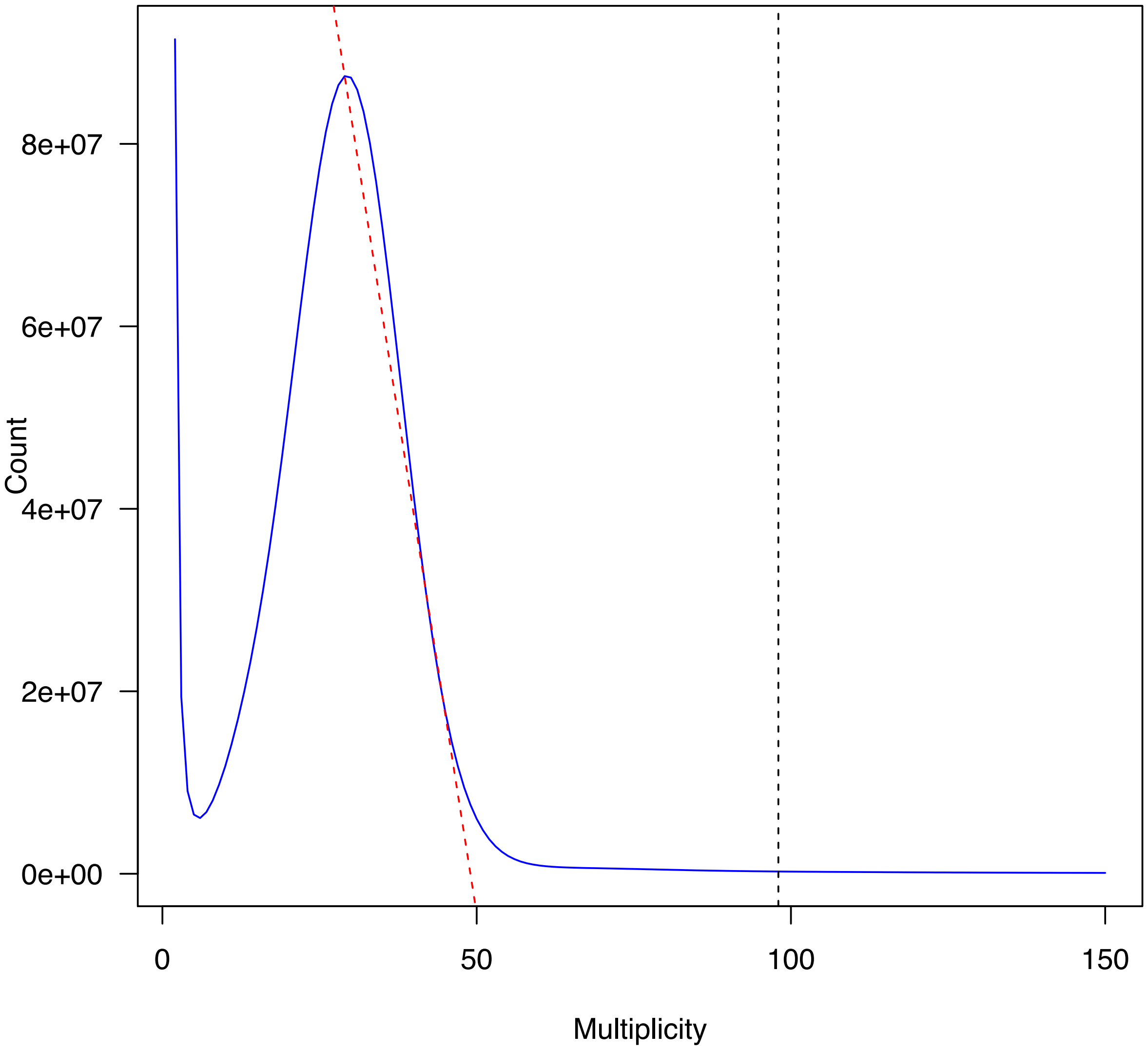


Figure S2. The count (A) and cumulative proportion (B) of unique 20-mers that are found *n* (multiplicity) times in the raw, paired-end sequencing reads (red line) and the trimmed and error-corrected paired-end reads (blue line). The vertical, dashed line indicates the threshold (*n* = 3) separating error *k*-mers (left of vertical line) from true *k*-mers (right of line).

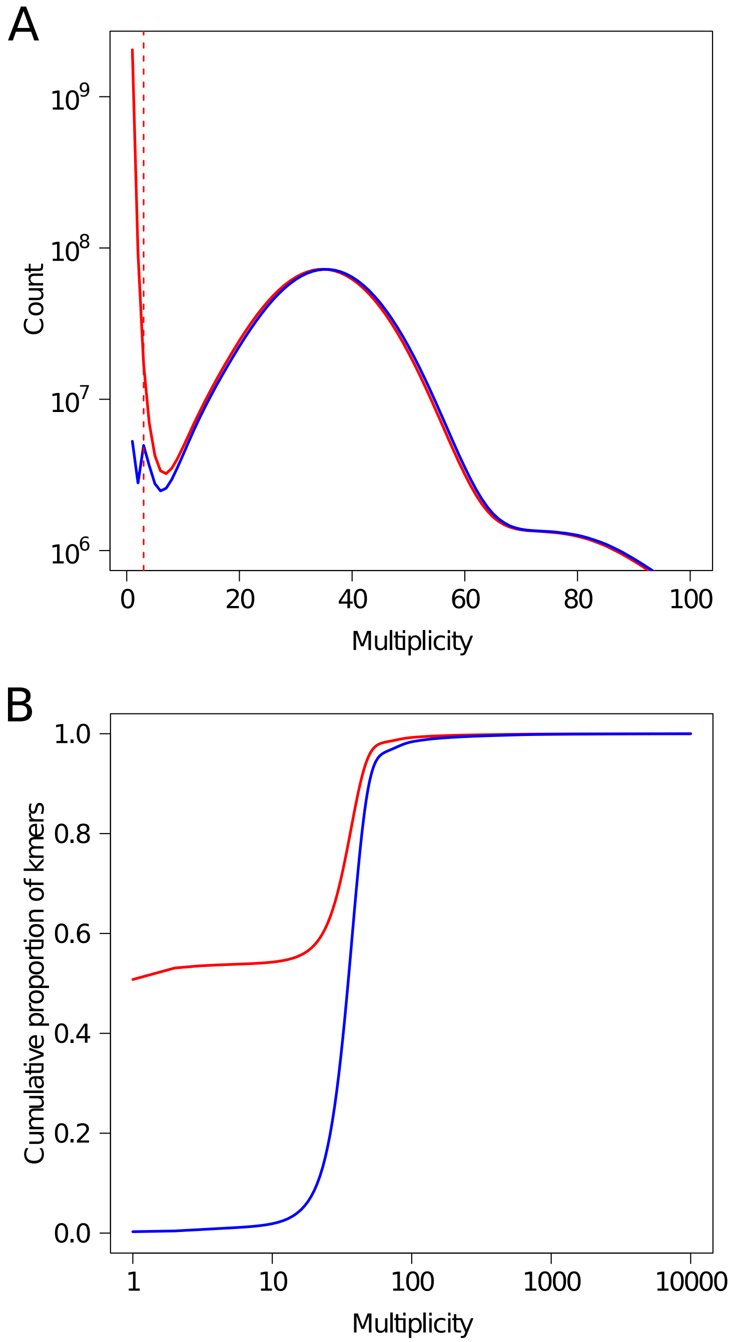


Figure S3. Histogram of the base quality scores corrected in the forward (green line), reverse (blue line) and unpaired (yellow line) raw reads.

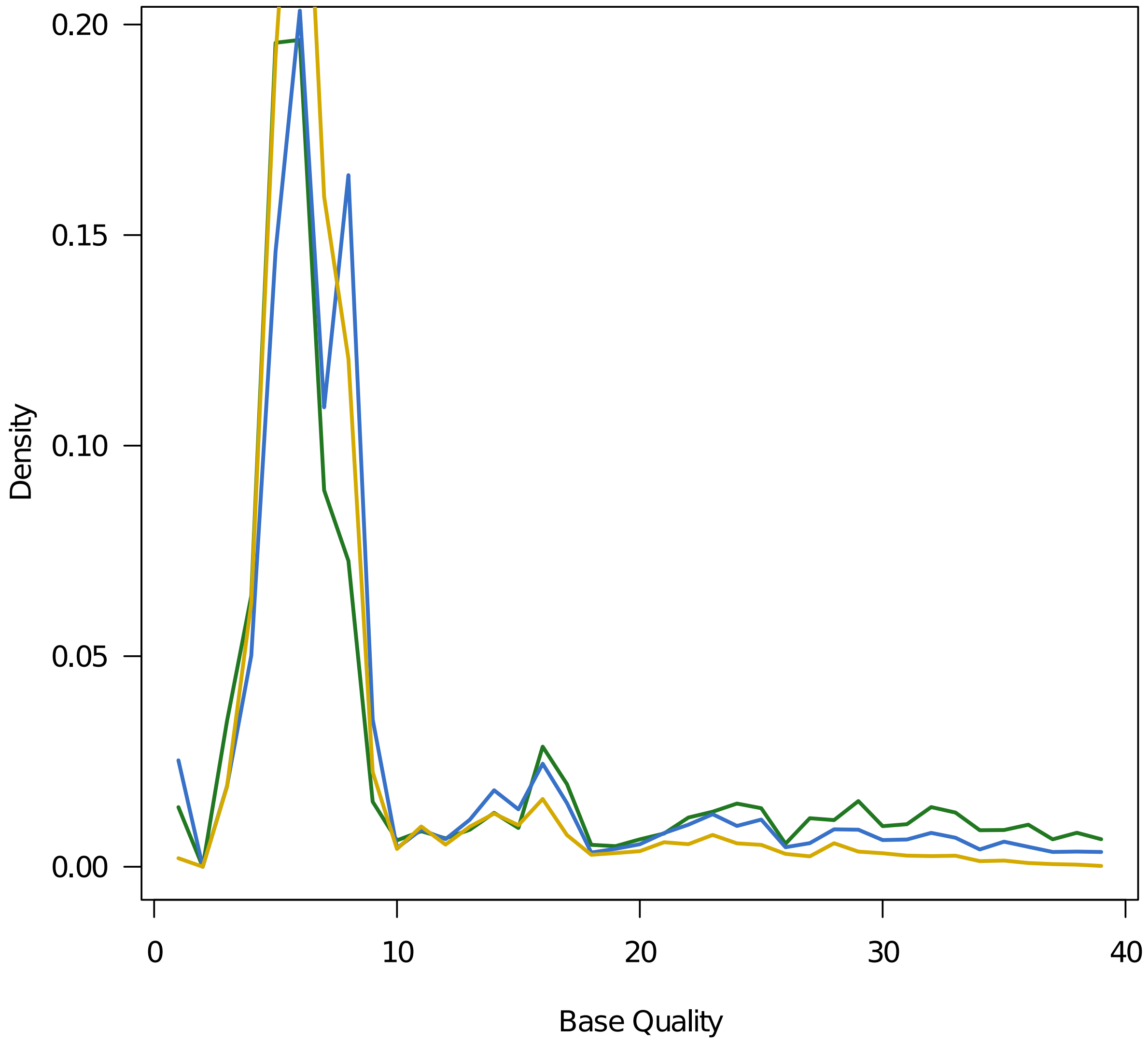


Figure S4. Comparison of (A) the number of scaffolds, (B) N50 length, (C) Proportion of 248 core eukaryotic genes (CEGs) annotated, and (D) longest scaffold length for various *k*-mer sizes used to assemble the genome. The trimmed and error-corrected paired-end reads were used for the assembly and trimmed mate-pair reads for the scaffolding step. Statistics are based upon scaffolds ≥500 bp in length.

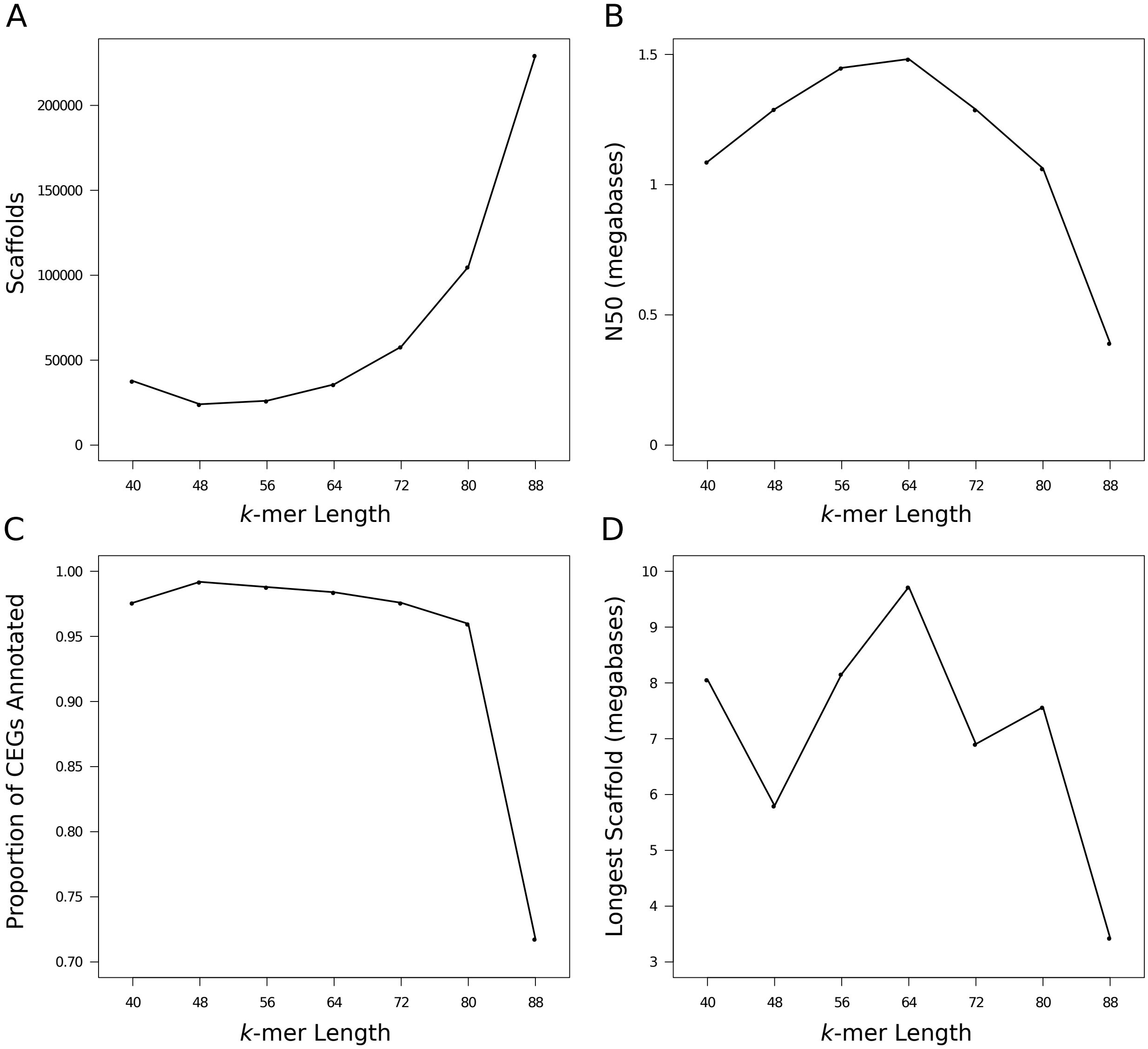


Figure S5. Distribution of the species (outer circle) and sequence types (inner circle) for the top blast hit for each of the short scaffolds (<500 bp) omitted from the final assembly. 1 = *C. dromedarius* microsatellite sequences, 2 = uncharacterized sequence clones, 3 = other, 4 = genes.

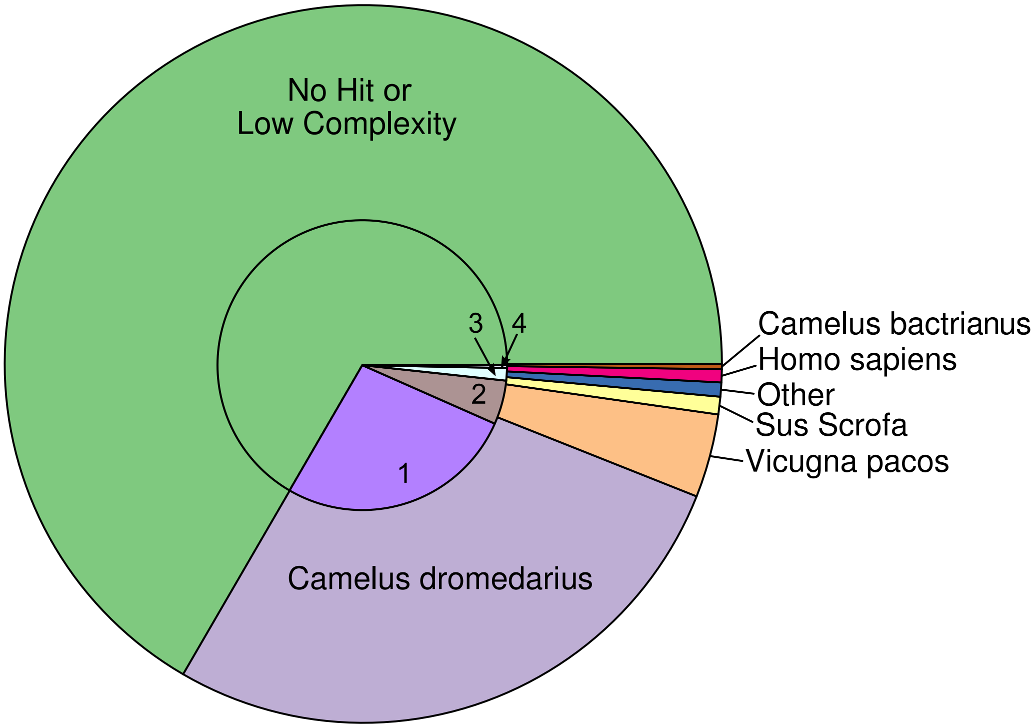


Figure S6. Histogram of the lengths (in base-pairs) of alignment blocks between our dromedary genome assembly and the reference (Accession no. GCA_000767585.1).

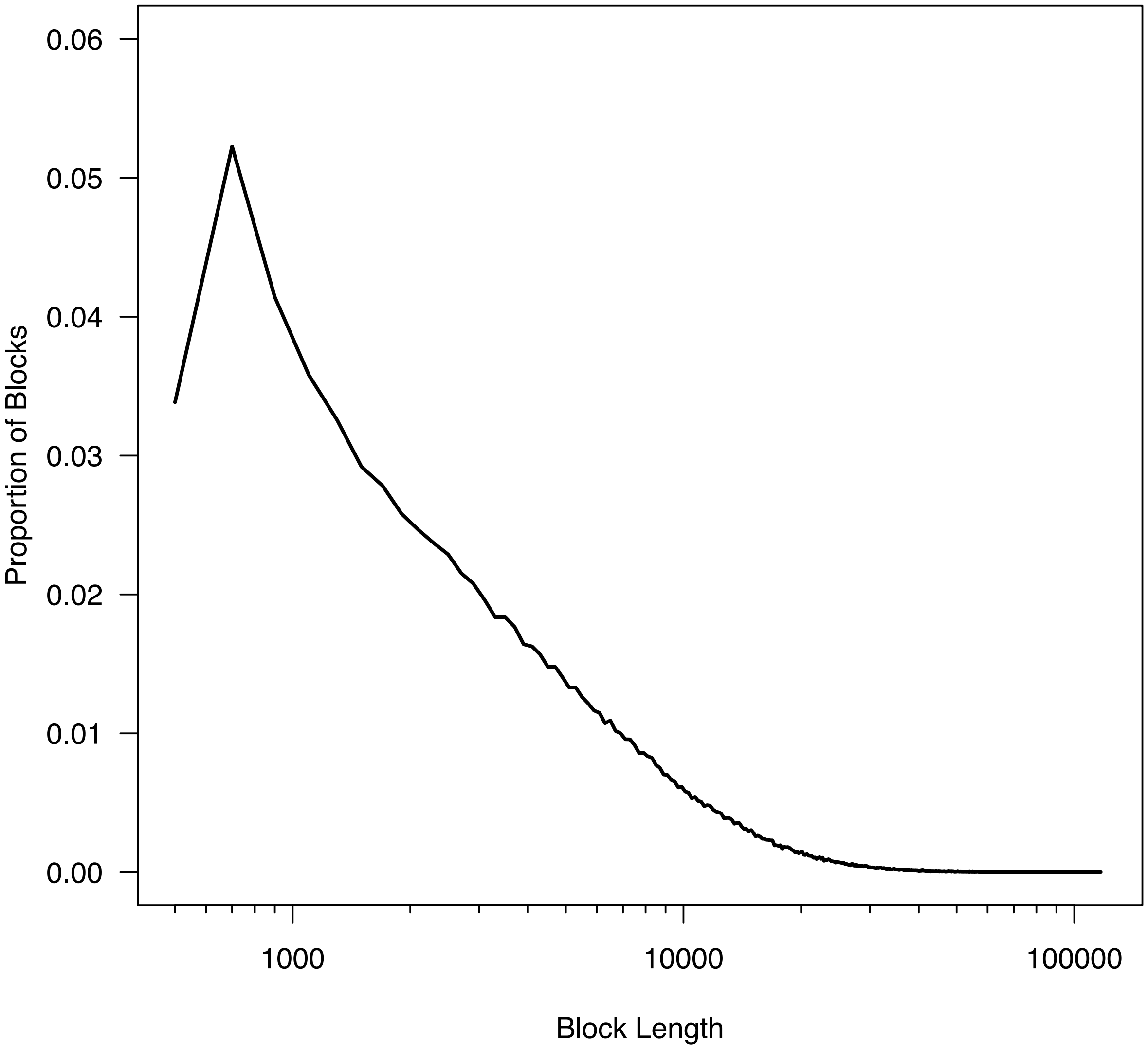


Figure S7. Cumulative number of genes ordered by increasing AED (annotation edit distance) scores. Only annotations with an AED score <0.75 (dashed line) were kept.

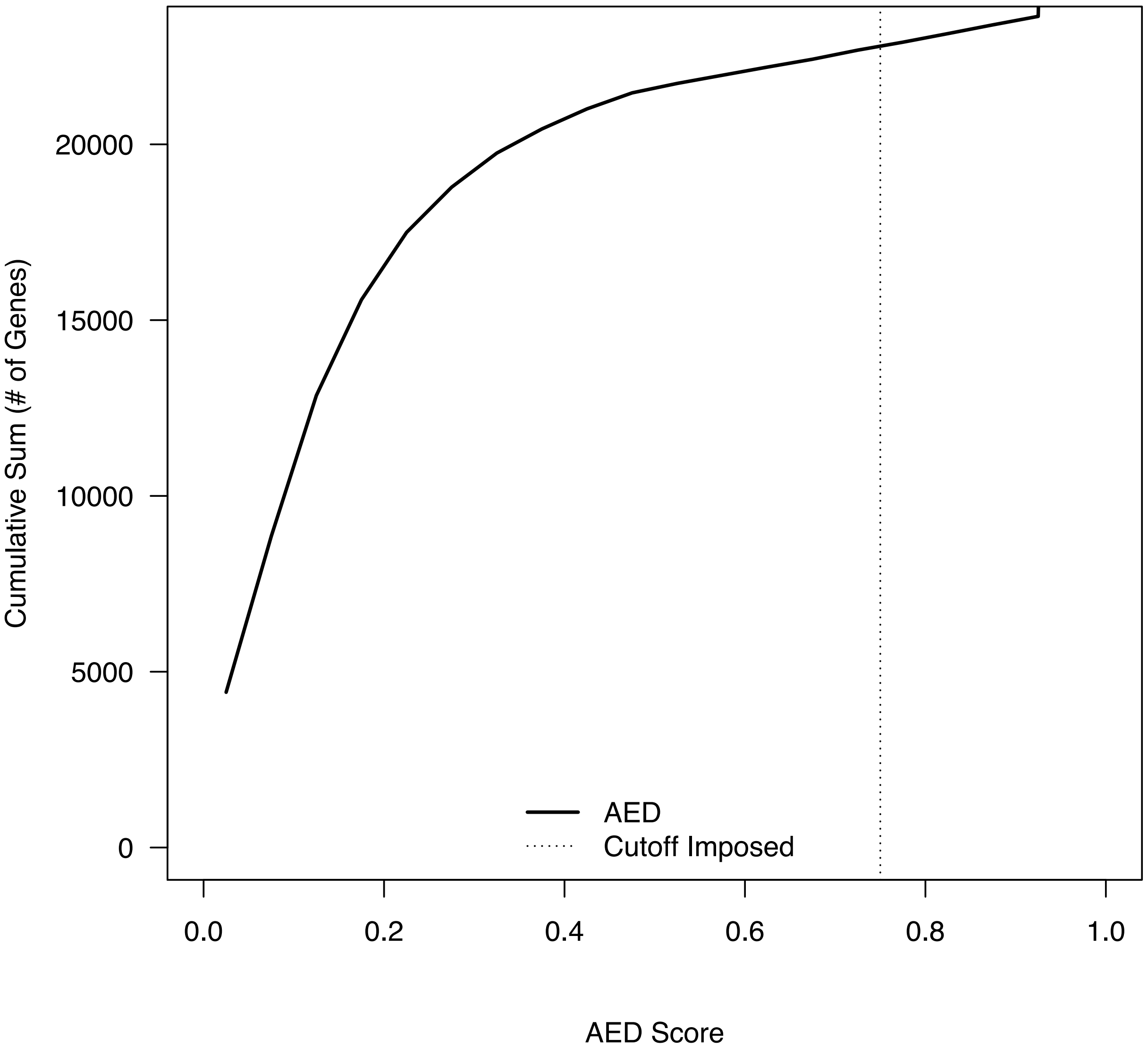


Figure S8. The distribution of (A) similarity scores for all the BLAST hits and (B) the species of the top hit for each annotated protein sequence. BLAST searches were performed against known metazoan protein sequences from Genbank's 'nr' database. Only the top 20 hits were kept for each gene with a minimum e-value of 10^{-3}. In (B), only the 25 most common species are shown.

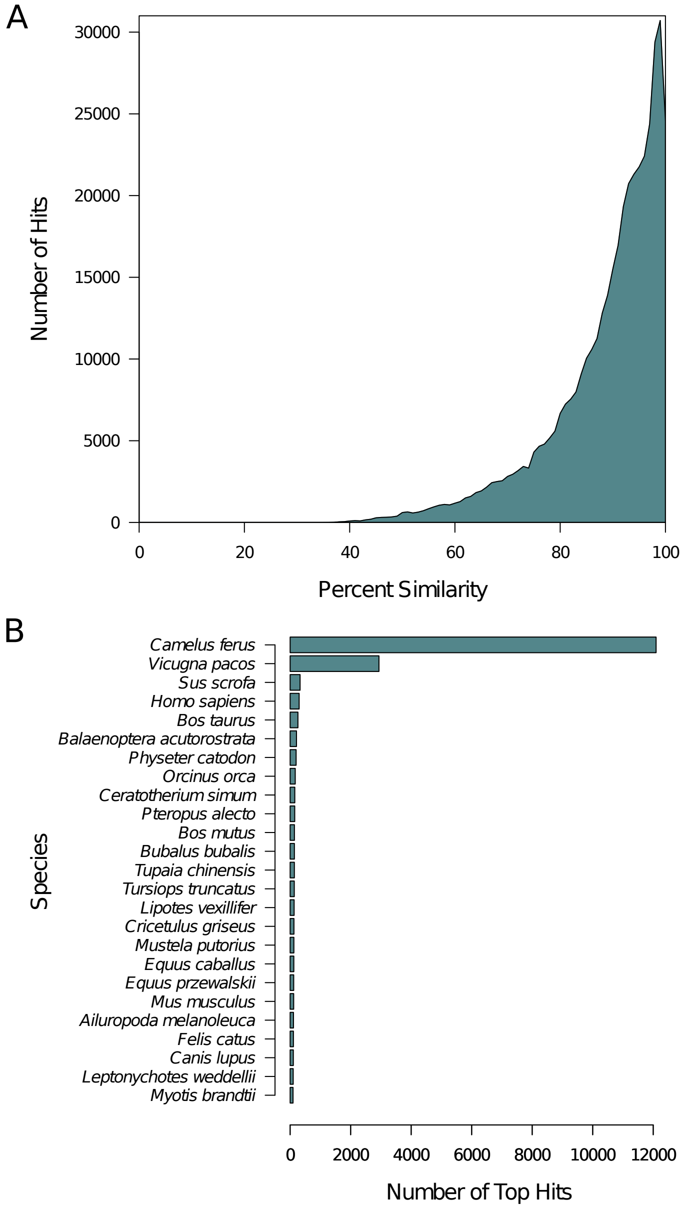


Figure S9. The number of Gene Ontology (GO) terms mapped to each protein sequence.

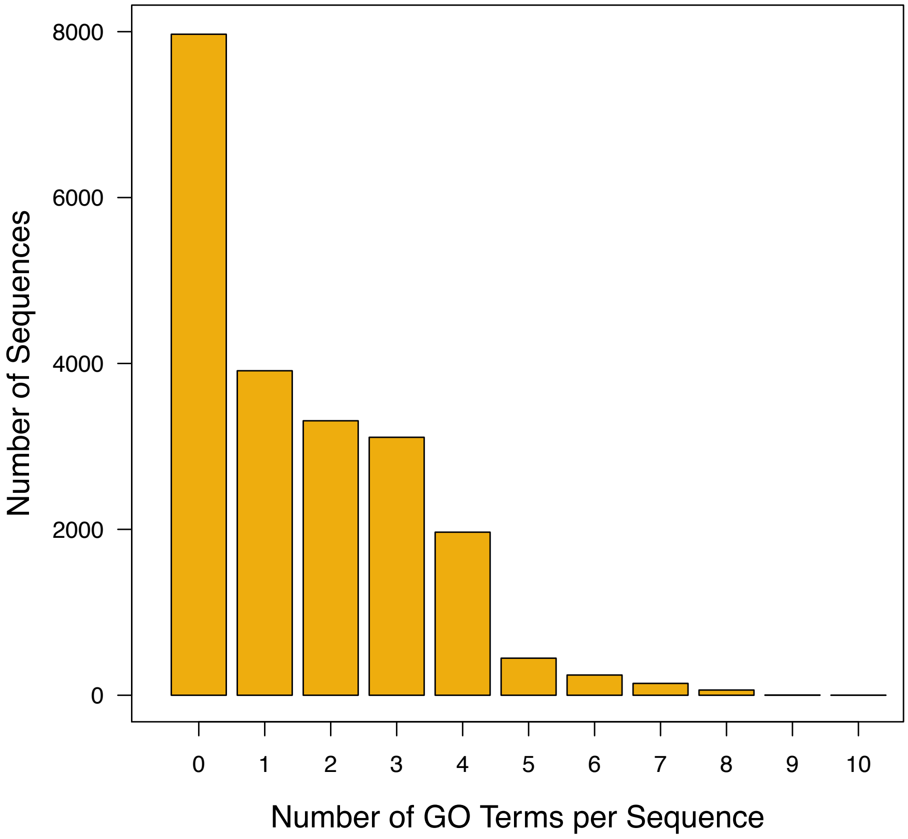


Figure S10. Histogram of the amino acid identity of single-copy orthologs between the African dromedary assembly and the *Camelus ferus* (solid line) and *Bos taurus* (dashed line) genome assemblies.

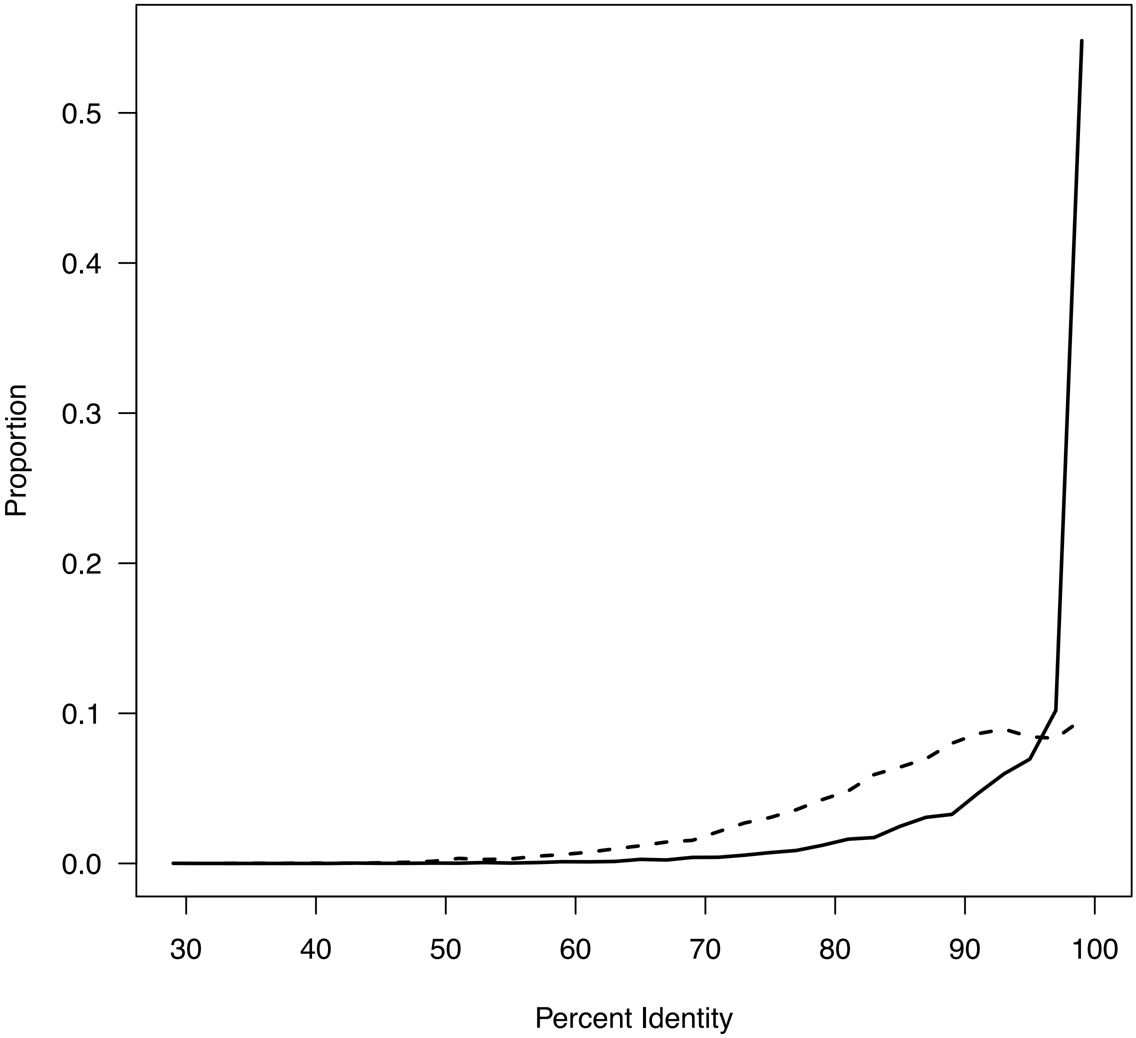
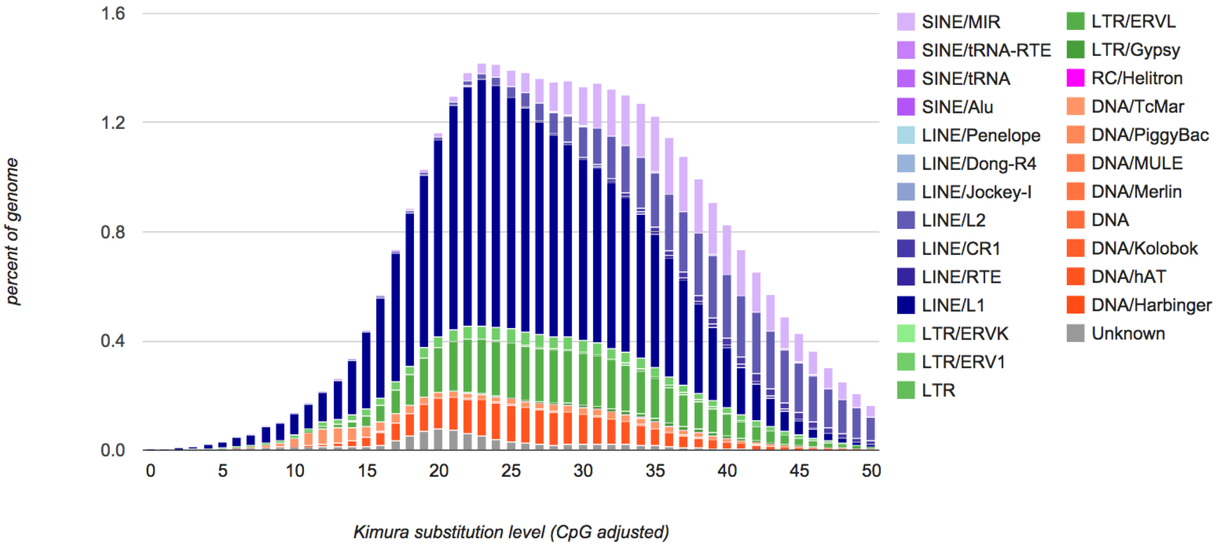


Figure S11. The relative abundance of repeat classes in the dromedary genome assembly versus the Kimura divergence from the consensus, using the combined set of annotated repetitive elements.



Methods S1. Example commands used for different analyses in this study. These are provided simply for reference and the reader should consult the software manuals for descriptions of the parameters used.

```
# Sequence trimming using POPOOLATION v1.2.2 for paired-end and mate-pair reads
trim-fastq.pl \
    –input Forward-PE-Reads.fq \
    –input Reverse-PE-Reads.fq \
    –quality-threshold 20 \
    –min-length 50
trim-fastq.pl \
        –input Forward-MP-Reads.fq \
        –input Reverse-MP-Reads.fq \
        –quality-threshold 20 \
        –min-length 30

# K-mer counting using DSK v1.6066 (k = 20)
dsk All-PE-Reads.fq 20 -t 1 -o dsk.k22.out -histo

# Error correction using QUAKE v0.3.5
parse_results dsk.k22.out.solid_kmers_binary > dsk_pe.k20.counts
cov_model.py --int dsk_pe.k20.counts
correct -f PE-reads.infile -k 20 -m dsk_pe.k20.counts -c 3 -p 16 –log

# Genome assembly using ABYSS v1.3.6 (e.g. k = 64)
abyss-pe \
        v=-v \
        np=16 \
        k=64 \
        n=5 \
        s=200 \
        name=Drom64K \
        lib=’drom’ \
        mp=’mp’ \
        drom=’Corrected-Forward-PE-Reads.fastq Corrected-Reverse-PE-Reads.fastq’ \
        se=’Corrected-SE-Reads.fq’ \
        mp=’ Forward-MP-Reads.fastq Reverse-MP-Reads.fastq’

# Assess core eukaryotic gene content using CEGMA v2.4.010312
cegma --mam -o Drom-CEGMA -v -T 16 -g Genome-assembly.fa

Basic quantitative analysis of an assembly using REAPR v1.0.16
smalt index \
        -k 13 -s 2 Genome-assembly_index Genome-assembly.fa
smalt sample \
    -u 1000 -n 16 -o Genome-assembly_sample \
    Genome-assembly_index Corrected-Forward-PE-Reads.fastq Corrected-Reverse-PE-Reads.fastq
smalt map \
    -r 0 -x -y 0.5 -n 16 -g Genome-assembly_sample \
    -f samsoft \
    Genome-assembly_index \
    Corrected-Forward-PE-Reads.fastq \
    Corrected-Reverse-PE-Reads.fastq | \
    awk '$1!~/^#/' | \
    samtools view -S -T Genome-assembly.fa -b - > smalt.raw.bam
samtools sort smalt.raw.bam smalt.raw.bam.sort
samtools rmdup smalt.raw.bam.sort.bam smalt.sort.rmdup.bam
echo "@HD VN:1.0 SO:coordinate" | awk '{OFS="\t"; $1=$1; print}' > smalt.header
samtools view -H smalt.sort.rmdup.bam >> smalt.header
samtools reheader smalt.header smalt.sort.rmdup.bam > smalt.final.bam
samtools index smalt.final.bam
reapr pipeline Genome-assembly.fa smalt.final.bam REAPR-OUT

# Align two dromedary genomes using MUGSY v1.2.3
mugsy -p Drom --directory . -d 500 -plot Cdrom-genbank.fa Genome-assembly.fa

# Calculate alignment statistics using MAFFILTER v1.1.0,
maffilter \
    input.file=Drom.maf \
    input.file.compression=none \
    output.log=Drom.maffilter.log \
    maf.filter=\
        AlnFilter(\
            species=(Cdrom_genbank,Drom64K_repmod),\
            window.size=10,\
            window.step=1,\
            max.gap=5,\
            missing_as_gap=yes,\
            file=data.trash_aln.maf.gz,\
            compression=gzip),\
        MinBlockSize(min_size=2),\
        MinBlockLength(min_length=500),\
        Output(\
            file=Drom.filtered.maf.gz,\
            compression=gzip,\
            mask=yes),\
        SequenceStatistics(\
                statistics=(\
                        BlockLength(),\
                        SequenceLength(\
                                species=Drom64K_repmod),\
                        SequenceLength(\
                                species=Cdrom_genbank),\
                        SiteStatistics(\
                                species=Drom64K_repmod),\
                        SiteStatistics(\
                                species=Cdrom_genbank),\
                        AlnScore(),\
                        PairwiseDivergence(\
                                species1=Cdrom_genbank,\
                                species2=Drom64K_repmod)),\
                ref_species=Cdrom_genbank,\
                file=Drom.filtered.statistics.blocks.csv),\
        WindowSplit(\
                preferred_size=500,\
                align=center),\
        SequenceStatistics(\
        statistics=(\
            BlockLength(),\
            SequenceLength(\
                species=Drom64K_repmod),\
            SequenceLength(\
                species=Cdrom_genbank),\
            SiteStatistics(\
                species=Drom64K_repmod),\
            SiteStatistics(\
                species=Cdrom_genbank),\
            AlnScore(),\
            PairwiseDivergence(\
                species1=Cdrom_genbank,\
                species2=Drom64K_repmod)),\
        ref_species=Drom64K_repmod,\
        file=Drom.filtered.statistics.windows.csv)

# Calculate alignment statistics using MAFFILTER v1.1.0, for features in a GFF file (e.g. CPG Islands)
maffilter \
    input.file=Drom.filtered.maf.gz \
    input.file.compression=gzip \
    output.log=Drom.maffilter.exons.log \
    maf.filter=\
        MinBlockSize(min_size=2),\
        ExtractFeature(\
            ref_species=Drom64K_repmod,\
            feature.file=exons.gff3,\
            feature.file.compression=none,\
            feature.format=GFF,\
            feature.type=all,\
            complete=yes,\
            ignore_strand=no),\
        SequenceStatistics(\
            statistics=(\
                BlockLength(),\
                SequenceLength(\
                    species=Drom64K_repmod),\
                SequenceLength(\
                    species=Cdrom_genbank),\
                SiteStatistics(\
                    species=Drom64K_repmod),\
                SiteStatistics(\
                    species=Cdrom_genbank),\
                AlnScore(),\
                PairwiseDivergence(\
                    species1=Cdrom_genbank,\
                    species2=Drom64K_repmod)),\
            ref_species=Drom64K_repmod,\
            file=Drom.statistics.cpg.csv),\
        Output(\
            file=Drom.cpg.maf.gz,\
            compression=gzip,\
            mask=yes)
```

```
# Extraction of divergent sites as a .vcf file from the genome alignment using MAFFILTER v1.1.0
maffilter \
        input.file=Drom.filtered.maf.gz \
        input.file.compression=gzip \
        output.log=Drom.maffilter.vcf.log \
        maf.filter=\
                VcfOutput(\
                        file=Drom.MAFsnp.vcf.gz,\
                        compression=gzip,\
                        reference=Drom64K_repmod,\
                        genotypes=(\
                                Drom64K_repmod,Cdrom_genbank))

# Predict genes using GENEMARK-ES
perl gm_es.pl -v Genome-assembly.fa

# Convert CEGMA results into SNAP hidden Markov model
cegma2zff Drom-CEGMA.cegma.gff Genome-assembly.fa
fathom genome.ann genome.dna -categorize 1000
fathom -export 1000 -plus uni.ann uni.dna
forge export.ann export.dna
hmm-assembler.pl drom . > cegmasnap.hmm

# Run first iteration of MAKER v2.31.6 # see Appendix 2A for MAKER configuration file
maker -base MAKER1

# Merge MAKER annotations into a single gff file and build new SNAP model
gff3_merge -d MAKER1.maker.output/MAKER1_master_datastore_index.log -o maker1_All.gff
maker2zff maker1_All.gff
fathom genome.ann genome.dna -categorize 1000
fathom -export 1000 -plus uni.ann uni.dna
forge export.ann export.dna
hmm-assembler.pl maker1 . > snap2.hmm

# Train a model for use with AUGUSTUS v2.5.5
autoAug.pl \
        --genome=Genome-assembly.fa \
        --species=dromedarius \
        --cdna=cDNA.fa \
        --trainingset=genome.gff3 \
        -v -v –v \
        –useexisting

# Run second iteration of MAKER v2.31.6 # see Appendix 2B for MAKER configuration files
maker -base MAKER2

# Assignment of orthologs using ORTHOMCL v2.0 (only comparison with Bos taurus is shown)
wget ftp://ftp.ensembl.org/pub/release-77/fasta/bos_taurus/pep/Bos_taurus.UMD3.1.pep.all.fa.gz
gunzip Bos_taurus.UMD3.1.pep.all.fa.gz
orthomclAdjustFasta Btau Bos_taurus.UMD3.1.pep.all.fa 1
orthomclAdjustFasta Cdro Drom.longestORFs.faa 1
orthomclFilterFasta . 10 20
makeblastdb -in goodProteins.fasta -dbtype prot -parse_seqids -out goodProteins.fasta
blastp -db goodProteins.fasta -query goodProteins.fasta -outfmt 6 -out blastresults.tsv
```

```
mkdir SEQS
mv Cdro.fasta SEQS
mv Btau.fasta SEQS
orthomclBlastParser blastresults.tsv ./SEQS >> similarSequences.txt
mysql -u rfitak –p
   DROP DATABASE orthomcl;
   create database orthomcl;
   exit
orthomclInstallSchema mysql.config mysql.log
orthomclLoadBlast mysql.config similarSequences.txt
orthomclPairs mysql.config pairs.log cleanup=no
orthomclDumpPairsFiles mysql.config
mcl mclInput --abc -I 1.5 -o groups_1.5.txt
orthomclMclToGroups OG1.5_ 1000 < groups_1.5.txt > named_groups_1.5.txt

# Masking the genome with REPEATMASKER v4.0.5
RepeatMasker –pa 16 –gff –xsmall Genome-assembly.fa

# Construct de novo repeat library with REPEATMODELER
# and mask the genome with REPEATMASKER
RepeatModeler/BuildDatabase –name Genome-assembly-db Genome-assembly.fa
RepeatModeler -engine ncbi -pa 15 -database Genome-assembly-db
RepeatMasker –pa 16 –gff –xsmall –lib consensi.fa.classified Genome-assembly.fa

# Constructing de novo repeat libraries from whole genome reads of dromedary genome
# using REPARK v1.2.2 and subsequent masking of the genome
# using REPEATMODELER and REPEATMASKER
perl RepARK.pl -l All-PE-Reads.fq -p 16 -d -o /RepARK/RepARK_working/
RepeatModeler/BuildDatabase –name Genome-RepArk-db  /RepARK/ velvet_repeat_lib/contigs.fa
RepeatModeler -engine ncbi -pa 15 -database Genome-RepArk-db
RepeatMasker –pa 16 –gff –xsmall –lib consensi.fa.classified Genome-assembly.fa

# Generate the repeat landscape using REPEATMASKER v4.0.5
perl \
    /RepeatMasker/util/calcDivergenceFromAlign.pl \
    -s Genome.fa.masked.cat. \
    divsum \
    Genome.fa.masked.cat.align.gz
perl \
        /RepeatMasker/util/createRepeatLandscape.pl \
        -div Genome.fa.masked.cat. \
        divsum > Genome_landscape.html

# Annotation of RNA sequences using INFERNAL v1.1 (only 1 Rfam model is shown,
# all families were searched)
GA=$(grep "^GA" RF00001.cm | sed 's/^GA[ ]*//g' | perl -ne '$_=0.85 * $_; print "$_"')
cmsearch \
        -Z 5400 \
        -T $GA \
        RF00001.cm \
        --tblout RNA.tbl \
        Genome-assembly.fa

# Identification of CpG islands using EMBOSS v6.5.7
```

```
perl -ne 'if ($_ =~ m/^>/){print "$_";}else{$_ =~ s/[acgt]/N/g; print "$_";}' \
    Drom64K_repmod.fa.masked > Drom64K.hardmasked.fasta
cpgplot \
            -sequence Drom64K.hardmasked.fasta \
            -outfile Drom64K.cpgplot.out \
            -noplot \
            -window 100 \
            -minlen 200 \
            -minoe 0.6 \
            -minpc 50 \
            -outfeat Drom64K.cpgplot.gff3

# Map the paired-end sequencing reads to the assembled genome using BWA 0.6.2
bwa aln \
    -n 0.01 \
    -o 1 \
    -e 12 \
    -d 12 \
    -l 32 \
    -t 16 \
    -I \
    Genome-assembly.fa \
    Corrected-Forward-PE-Reads.fastq > Fwd.sai
bwa aln \
    -n 0.01 \
    -o 1 \
    -e 12 \
    -d 12 \
    -l 32 \
    -t 16 \
    -I \
    Genome-assembly.fa \
    Corrected-Reverse-PE-Reads.fastq > Rev.sai
bwa sampe \
            –r $rg \
            Genome-assembly.fa \
            Fwd.sai \
            Rev.sai \
            Corrected-Forward-PE-Reads.fastq \
            Corrected-Reverse-PE-Reads.fastq | \
            samtools view -u - > Dromedary.bam

# Convert alignments to sorted ‘bam’ format and filter for high-quality, properly paired reads
# using SAMTOOLS v1.1 (‘rmdup’ used SAMTOOLS v0.1.19)
samtools view \
            -u \
            -q 20 \
            -f 0x0002 \
            -F 0x0004 \
            -F 0x0008 \
            Dromedary.bam | \
            samtools rmdup - - | \
            samtools sort \
            -O bam \
            -T Drom.sorted - > Drom.sorted.rmdup.mq20.bam

# Call SNPs using SAMTOOLS v1.1
samtools mpileup \
            -C50 \
            -t DP,DPR,DV,DP4,INFO/DPR,SP \
            -uf Genome-assembly.fa \
            Drom.sorted.rmdup.mq20.bam | \
            bcftools call \
            -O v \
            -c -M \
            -A \
            -v - > Drom.samtools.raw.vcf

# Call SNPs using PLATYPUS v0.7.9.1
Platypus.py callVariants \
            -o Drom.platypus.raw.vcf \
            --refFile=Genome-assembly.fa \
            --bamFiles=Drom.sorted.rmdup.mq20.bam \
            --nCPU 16

# Reduce MNVs to SNVs using GATK VariantsToAllelicPrimitives
java -Xmx20g -jar GenomeAnalysisTK.jar \
            -T VariantsToAllelicPrimitives \
            -R Genome-assembly.fa \
            --variant Drom.platypus.raw.vcf \
            -o Drom.platypus.primitives.vcf

# Keep only SNPs with "PASS" from PLATYPUS variants using VCFTOOLS v.0.1.12
vcftools \
            --vcf Drom.platypus.primitives.vcf \
            --remove-indels \
            --recode \
            --recode-INFO-all \
            --remove-filtered-all \
            --out Drom.platypus.primitives.filtered.vcf

# Find the intersection of raw SAMTOOLS SNPs with the filtered PLATYPUS SNPs
# using BEDTOOLS v.2.17.0
intersectBed \
            -wa \
            -header \
            -a Drom.samtools.rawSNPs.vcf \
            -b Drom.platypus.primitives.filtered.vcf > Drom-SAM-PL.overlap.vcf

# Filter and annotate the overlapping SNPs
# using SAMTOOLS/BCFTOOLS v1.1 and VCFTOOLS v.0.1.12
bcftools filter \
    -O v \
    -g5 \
    -G5 \
    -i 'QUAL>=20 && DP>=14 && DP<=86' \
    -s QUAL-DP < Drom-SAM-PL.overlap.vcf | \
    bcftools filter \
    -O v \
    -i 'INFO/PLATYPUS!="FAIL"' \
    -s PLATYPUS \
    -m+ - > Drom.overlap.filtered.vcf
vcftools \
        --vcf Drom.overlap.filtered.vcf \
        --recode \
        --recode-INFO-all \
        --remove-filtered-all \
        --out Drom.final.vcf

# Calculate Ti/Tv ratio using VCFTOOLS v.0.1.12
vcftools \
        --vcf Drom.final.vcf \
        --remove-indels \
        --TsTv-summary \
        --out Drom.TiTv

# Calculate SNP density in 1kb windows using VCFTOOLS v.0.1.12
vcftools \
        --vcf Drom.final.vcf \
        --remove-indels \
        --SNPdensity 1000 \
        --out Drom

# Calculate SNP density in annotated regions using VCFTOOLS v.0.1.12 (only exons shown)
vcftools \
        --vcf Drom.final.vcf \
        --remove-indels \
        --bed exons.bed \
        --recode \
        --out Drom.exons

# Demographic history using PSMC and the default parameters,
# with a coverage cutoff minimum of 14x
# and maximum of 86x (1/3 and 2x the mean coverage).
samtools mpileup \
    -C50 \
    -S \
    -D \
    -uf Genome-assembly.fa \
    Drom.sorted.rmdup.mq20.bam | \
    bcftools view \
    -c - | \
    vcfutils.pl vcf2fq \
    -d 14 \
    -D 86 | \
    gzip > Drom.raw.fq.gz
fq2psmcfa Drom.raw.fq.gz >  Drom.raw.psmcfa
splitfa Drom.raw.psmcfa > Drom.raw.split.psmcfa
psmc \
    -N25 \
    -t15 \
    -r5 \
        -p "4+25*2+4+6" \
        Drom.raw.psmcfa \
        -o Drom.raw.psmc
for i in {1..100}; \
        do \
                psmc \
                -N25 \
                -t15 \
                -r5 \
                -b \
                -p "4+25*2+4+6" \
                -o Drom.raw.$i.psmc \
                Drom.raw.split.psmcfa; \
        done
cat \
        Drom.raw.psmc \
        Drom.raw.*.psmc > bootstrapped.raw.psmc
psmc_plot.pl \
                -P bottom \
                -X2000000 -p \
                -g5 \
                -x1000 \
                Plot.bootstrapped.raw \
                bootstrapped.raw.psmc

# Demographic history using PSMC with the repeat-masked genome and the filtered set of SNPs.
bcftools consensus \
        -f Genome.fa.masked \
        -i Drom.overlap.filtered.vcf > Drom.filtered.fa
fq2psmcfa Drom.filtered.fa > Drom.filtered.psmcfa
splitfa Drom.filtered.psmcfa > Drom.filtered.split.psmcfa
psmc \
        -N25 \
        -t15 \
        -r5 \
        -p "4+25*2+4+6" \
        Drom.filtered.psmcfa -o Drom.filtered.psmc
for i in {1..100}; \
        do \
                psmc \
                -N25 \
                -t15 \
                -r5 \
                -b \
                -p "4+25*2+4+6" \
                -o Drom.filtered.$i.psmc \
                Drom.filtered.split.psmcfa; \
        done
psmc_plot.pl \
                -P bottom \
                -X2000000 \
                -p \
                -g5 \
                -x1000 \
                Plot.bootstrapped.filtered \
                bootstrapped.filtered.psmc
```

Methods S2. Configuration files for the first (A) and second (B) iterations of MAKER v2.31.6

A. maker_opts.ctl configuration file for the first iteration of MAKER

```
#-----Genome (these are always required)
genome=Genome-assembly.fa #genome sequence (fasta file or fasta embeded in GFF3 file)
organism_type=eukaryotic #eukaryotic or prokaryotic. Default is eukaryotic
#-----Re-annotation Using MAKER Derived GFF3
maker_gff= #MAKER derived GFF3 file
est_pass=0 #use ESTs in maker_gff: 1 = yes, 0 = no
altest_pass=0 #use alternate organism ESTs in maker_gff: 1 = yes, 0 = no
protein_pass=0 #use protein alignments in maker_gff: 1 = yes, 0 = no
rm_pass=0 #use repeats in maker_gff: 1 = yes, 0 = no
model_pass=0 #use gene models in maker_gff: 1 = yes, 0 = no
pred_pass=0 #use ab-initio predictions in maker_gff: 1 = yes, 0 = no
other_pass=0 #passthrough anyything else in maker_gff: 1 = yes, 0 = no
#-----EST Evidence (for best results provide a file for at least one)
est=cDNA.fa #set of ESTs or assembled mRNA-seq in fasta format
altest= #EST/cDNA sequence file in fasta format from an alternate organism
est_gff= #aligned ESTs or mRNA-seq from an external GFF3 file
altest_gff= #aligned ESTs from a closly relate species in GFF3 format
#-----Protein Homology Evidence (for best results provide a file for at least one)
protein=homologous-proteins.fa  #protein sequence file in fasta format (i.e. from cow, Bactrian camel, and alpaca)
protein_gff=  #aligned protein homology evidence from an external GFF3 file
#-----Repeat Masking (leave values blank to skip repeat masking)
model_org=all #select a model organism for RepBase masking in RepeatMasker
rmlib= #provide an organism specific repeat library in fasta format for RepeatMasker
repeat_protein=te_proteins.fasta #provide a fasta file of transposable element proteins for RepeatRunner
rm_gff= #pre-identified repeat elements from an external GFF3 file
prok_rm=0 #forces MAKER to repeatmask prokaryotes (no reason to change this), 1 = yes, 0 = no
softmask=1 #use soft-masking rather than hard-masking in BLAST (i.e. seg and dust filtering)
#-----Gene Prediction
snaphmm=cegmasnap.hmm #SNAP HMM file
gmhmm=es.mod #GeneMark HMM file
augustus_species= #Augustus gene prediction species model
fgenesh_par_file= #FGENESH parameter file
pred_gff= #ab-initio predictions from an external GFF3 file
model_gff= #annotated gene models from an external GFF3 file (annotation pass-through)
est2genome=1 #infer gene predictions directly from ESTs, 1 = yes, 0 = no
protein2genome=1 #infer predictions from protein homology, 1 = yes, 0 = no
trna=1 #find tRNAs with tRNAscan, 1 = yes, 0 = no
snoscan_rrna= #rRNA file to have Snoscan find snoRNAs
unmask=0 #also run ab-initio prediction programs on unmasked sequence, 1 = yes, 0 = no
#-----Other Annotation Feature Types (features MAKER doesn't recognize)
other_gff= #extra features to pass-through to final MAKER generated GFF3 file
#-----External Application Behavior Options
alt_peptide=C #amino acid used to replace non-standard amino acids in BLAST databases
cpus=1 #max number of cpus to use in BLAST and RepeatMasker (not for MPI, leave 1 when using MPI)
#-----MAKER Behavior Options
max_dna_len=100000 #length for dividing up contigs into chunks (increases/decreases memory usage)
min_contig=1 #skip genome contigs below this length (under 10kb are often useless)
pred_flank=200 #flank for extending evidence clusters sent to gene predictors
pred_stats=0 #report AED and QI statistics for all predictions as well as models
```

```
AED_threshold=1 #Maximum Annotation Edit Distance allowed (bound by 0 and 1)
min_protein=0 #require at least this many amino acids in predicted proteins
alt_splice=0 #Take extra steps to try and find alternative splicing, 1 = yes, 0 = no
always_complete=0 #extra steps to force start and stop codons, 1 = yes, 0 = no
map_forward=0 #map names and attributes forward from old GFF3 genes, 1 = yes, 0 = no
keep_preds=1 #Concordance threshold to add unsupported gene prediction (bound by 0 and 1)
split_hit=10000 #length for the splitting of hits (expected max intron size for evidence alignments)
single_exon=1 #consider single exon EST evidence when generating annotations, 1 = yes, 0 = no
single_length=250 #min length required for single exon ESTs if 'single_exon is enabled'
correct_est_fusion=0 #limits use of ESTs in annotation to avoid fusion genes
tries=2 #number of times to try a contig if there is a failure for some reason
clean_try=0 #remove all data from previous run before retrying, 1 = yes, 0 = no
clean_up=0 #removes theVoid directory with individual analysis files, 1 = yes, 0 = no
TMP= #specify a directory other than the system default temporary directory for temporary files
```

B. maker_opts.ctl configuration file for the second iteration of MAKER

```
#-----Genome (these are always required)
genome=Genome-assembly.fa #genome sequence (fasta file or fasta embeded in GFF3 file)
organism_type=eukaryotic #eukaryotic or prokaryotic. Default is eukaryotic
#-----Re-annotation Using MAKER Derived GFF3
maker_gff= #MAKER derived GFF3 file
est_pass=0 #use ESTs in maker_gff: 1 = yes, 0 = no
altest_pass=0 #use alternate organism ESTs in maker_gff: 1 = yes, 0 = no
protein_pass=0 #use protein alignments in maker_gff: 1 = yes, 0 = no
rm_pass=0 #use repeats in maker_gff: 1 = yes, 0 = no
model_pass=0 #use gene models in maker_gff: 1 = yes, 0 = no
pred_pass=0 #use ab-initio predictions in maker_gff: 1 = yes, 0 = no
other_pass=0 #passthrough anyything else in maker_gff: 1 = yes, 0 = no
#-----EST Evidence (for best results provide a file for at least one)
est=cDNA.fa #set of ESTs or assembled mRNA-seq in fasta format
altest= #EST/cDNA sequence file in fasta format from an alternate organism
est_gff= #aligned ESTs or mRNA-seq from an external GFF3 file
altest_gff= #aligned ESTs from a closly relate species in GFF3 format
#-----Protein Homology Evidence (for best results provide a file for at least one)
protein=homologous-proteins.fa  #protein sequence file in fasta format (i.e. from cow, Bactrian camel,
        alpaca)
protein_gff=  #aligned protein homology evidence from an external GFF3 file
#-----Repeat Masking (leave values blank to skip repeat masking)
model_org=all #select a model organism for RepBase masking in RepeatMasker
rmlib= #provide an organism specific repeat library in fasta format for RepeatMasker
repeat_protein=te_proteins.fasta #provide a fasta file of transposable element proteins for RepeatRunner
rm_gff= #pre-identified repeat elements from an external GFF3 file
prok_rm=0 #forces MAKER to repeatmask prokaryotes (no reason to change this), 1 = yes, 0 = no
softmask=1 #use soft-masking rather than hard-masking in BLAST (i.e. seg and dust filtering)
#-----Gene Prediction
snaphmm=snap2.hmm #SNAP HMM file
gmhmm=es.mod #GeneMark HMM file
augustus_species=dromedarius #Augustus gene prediction species model
fgenesh_par_file= #FGENESH parameter file
pred_gff= #ab-initio predictions from an external GFF3 file
model_gff= #annotated gene models from an external GFF3 file (annotation pass-through)
est2genome=0 #infer gene predictions directly from ESTs, 1 = yes, 0 = no
protein2genome=0 #infer predictions from protein homology, 1 = yes, 0 = no
trna=1 #find tRNAs with tRNAscan, 1 = yes, 0 = no
```

```
snoscan_rrna= #rRNA file to have Snoscan find snoRNAs
unmask=0 #also run ab-initio prediction programs on unmasked sequence, 1 = yes, 0 = no
#-----Other Annotation Feature Types (features MAKER doesn't recognize)
other_gff= #extra features to pass-through to final MAKER generated GFF3 file
#-----External Application Behavior Options
alt_peptide=C #amino acid used to replace non-standard amino acids in BLAST databases
cpus=1 #max number of cpus to use in BLAST and RepeatMasker (not for MPI, leave 1 when using MPI)
#-----MAKER Behavior Options
max_dna_len=100000 #length for dividing up contigs into chunks (increases/decreases memory usage)
min_contig=1 #skip genome contigs below this length (under 10kb are often useless)

pred_flank=200 #flank for extending evidence clusters sent to gene predictors
pred_stats=1 #report AED and QI statistics for all predictions as well as models
AED_threshold=1 #Maximum Annotation Edit Distance allowed (bound by 0 and 1)
min_protein=30 #require at least this many amino acids in predicted proteins
alt_splice=1 #Take extra steps to try and find alternative splicing, 1 = yes, 0 = no
always_complete=0 #extra steps to force start and stop codons, 1 = yes, 0 = no
map_forward=0 #map names and attributes forward from old GFF3 genes, 1 = yes, 0 = no
keep_preds=1 #Concordance threshold to add unsupported gene prediction (bound by 0 and 1)
split_hit=10000 #length for the splitting of hits (expected max intron size for evidence alignments)
single_exon=1 #consider single exon EST evidence when generating annotations, 1 = yes, 0 = no
single_length=250 #min length required for single exon ESTs if 'single_exon is enabled'
correct_est_fusion=0 #limits use of ESTs in annotation to avoid fusion genes
tries=2 #number of times to try a contig if there is a failure for some reason
clean_try=0 #remove all data from previous run before retrying, 1 = yes, 0 = no
clean_up=0 #removes theVoid directory with individual analysis files, 1 = yes, 0 = no
TMP= #specify a directory other than the system default temporary directory for temporary files
```