Genetic diversity in the lion (panthera leo (Linnaeus 1758)) : unravelling the past and prospects for the future

Bertola, L.D.

Chapter 5

SNP discovery and phylogenetic analyses across ten populations of lions reveals a more complex evolutionary history

(in prep.)

L. D. Bertola, M. Vermaat, P. A. White, H. H. de Iongh, J. Laros and K. Vrieling

Abstract

Next generation sequencing techniques allow for the generation of new magnitudes of unlinked genetic markers, which can be used to infer phylogeographic patterns in non-model organisms, such as the lion. Previous studies in lions, mostly based on mtDNA and microsatellite markers, have shown that the distribution of genetic diversity is not in line with the current taxonomy, only distinguishing an African and an Asiatic subspecies. The addition of genome-wide, unlinked genetic markers provides us with a more complete picture of the underlying genomic complexity. Full genome sequencing and subsequent variant calling has resulted in the discovery of 44,627 SNPs in ten lions, sampled throughout their geographic range, one leopard and one tiger. A total of 18,457 SNPs was variable within the lion. Phylogenetic trees based on autosomal SNPs show a gradual differentiation in the lion, following a north-south axis, and no reciprocally monophyletic groups could be identified. However, the Asiatic subspecies shows a nested position within the African subspecies, indicating that the current nomenclature does not follow the deepest evolutionary split for the distinction of subspecies. Phylogenetic trees based on the mitochondrial genome show a strongly supported split between lions from the northern part of their range, and lions from the southern part of their range. Since autosomal SNP data do not show a conflicting pattern, we suggest that this distinction should be followed in a taxonomic revision of the lion.

Keywords: Single Nucleotide Polymorphism (SNP), SNP discovery, genome sequencing, phylogeography, lion (*Panthera leo*)

Introduction

The rise of next generation sequencing (NGS) techniques has opened up possibilities to apply massive parallel sequencing to non-model organisms, like the lion (*Panthera leo*). Inferring population histories and reconstruction of the evolutionary history of a species can therefore be based on a new magnitude of unlinked data. Species histories are favorably based on data from multiple loci, due to the fact that genetic markers may represent different evolutionary trajectories (mtDNA vs. autosomal DNA) and due to stochasticity in the coalescence of markers (Edwards 2009; Knowles 2009). Mitochondrial DNA (mtDNA) has been proven to be a useful tool for gaining insight in phylogeographic patterns, partially because of its shorter coalescence time compared to nuclear markers. However, it represents one locus only and obtained haplotype trees are therefore not necessarily a true representation of the underlying genomic complexity (Zink & Barrowclough 2008; Edwards & Bensch 2009).

The lion (*Panthera leo*) was subjected to several phylogeographic studies which have contributed to current insights into the distribution of genetic diversity in the African subspecies (*Panther leo leo*) and its connection to the Asiatic subspecies (*Panthera leo persica*). These studies included data from mtDNA (Dubach *et al.* 2005, 2013; Barnett *et al.* 2006a; b, 2014; Antunes *et al.* 2008; Bertola *et al.* 2011; Bertola *et al.* submitted), autosomal DNA (Antunes *et al.* 2008; Dubach *et al.* 2013; Bertola *et al.* submitted) and pathogens (Antunes *et al.* 2008). The general emerging pattern was that of a basal dichotomy, recognizing a Northern group with populations from West and Central Africa including the Asiatic subspecies (*Panthera leo persica*), and a Southern group with populations from East and Southern Africa (Bertola *et al.* 2011a; Dubach *et al.* 2013; Barnett *et al.* 2014). Within these two groups, more phylogenetic lineages can be recognized, with notably long lineages in the Southern group. Admixture between haplogroups was only found in two occasions: 1) haplotypes from both the Central and the North East group are found in the suture zone in Ethiopia, and 2) haplotypes from the South West group and the East/Southern are found in the Kruger/Limpopo area, Republic of South Africa (RSA), likely to be the result of human-mediated translocations (Bertola *et al*. submitted) (Miller *et al.* 2013) (Figure 1). Microsatellite data are roughly congruent, also identifying a distinct position for the West and Central African lion, and a subsequent split between East and Southern Africa in the populations from the southern part of the range. However, there is a geographic discrepancy in the southern part of the range, where admixture is clearly visible, notably between East/Southern and South West Africa based on autosomal data only (Figure 1). The above mentioned studies have not only given a fine scale picture of current genetic diversity in the lion, but also illustrate how this diversity deviates from the current taxonomic nomenclature, only recognizing an African and an Asiatic subspecies (Bertola *et al.* 2011a; Dubach *et al.* 2013; Barnett *et al.* 2014) (Bertola *et al.,* submitted). This has led to requests for a taxonomic revision for this species (Bertola *et al.* 2011a; Dubach *et al.* 2013; Barnett *et al.* 2014) (Bertola *et al.,* submitted)

Although mtDNA and autosomal data have not shown strongly conflicting patterns in lion phylogeography, additional data from genome wide markers would benefit both the understanding of the evolutionary history of the species, and guiding of conservation efforts. According to Moritz (1994) intraspecific genetic diversity can be used as a rational for conservation practices, by following a two-step approach and defining 1) Evolutionary Significant Units (ESUs), and 2) Management Units (MUs). The inclusion of nuclear data for the recognition of ESUs is essential to avoid misclassifying

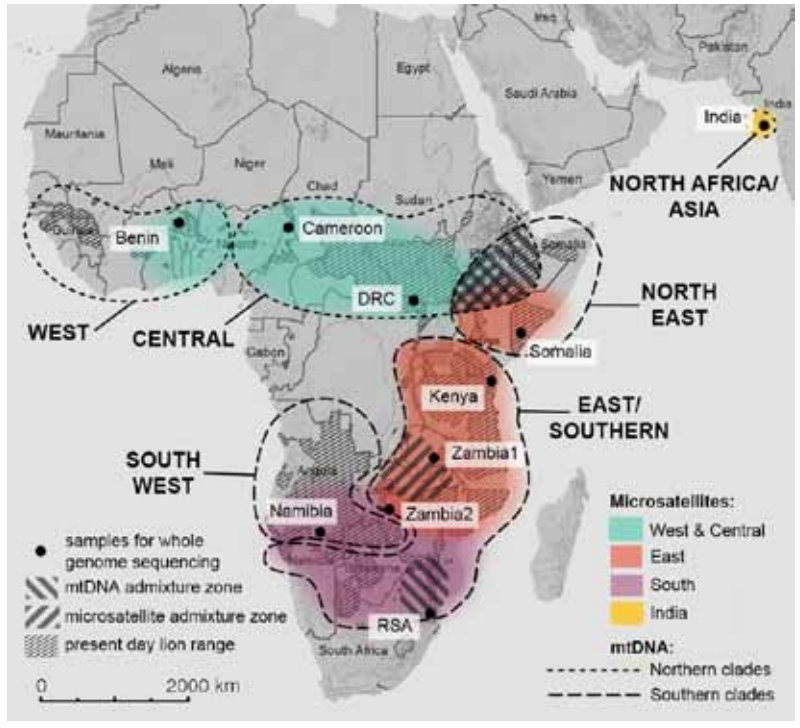


Figure 1. Phylogeographic groups in the lion identified based on mtDNA (Bertola *et al.*, submitted) and microsatellite data (Bertola *et al.*, submitted), and locations of lion samples included in this study. Lion range data from IUCN (2014).

populations which are linked by nuclear, but not by organellar gene flow (Moritz 1994). It is known that female lions exhibit strong philopatry and that male lions are capable dispersers (Pusey *et al.* 1987; Spong & Creel 2001), indicating that this aspect may be relevant in this species. In previous studies on lions which contained autosomal data, few populations were included as representatives. Also autosomal data were mainly represented by microsatellite loci (Antunes *et al.* 2008; Dubach *et al.* 2013; Bertola *et al.* submitted), which are of limited use to infer phylogenetic relationships due to their mutation pattern and high variability. To derive a complete picture of the evolutionary history, a broader range of autosomal markers should be targeted, and compared to the available mtDNA datasets.

In this study, we describe the discovery of Single Nucleotide Polymorphisms (SNPs) by targeting variable positions from whole genome data of ten lions, covering the main phylogeographic groups as were indicated based on previously published mtDNA and microsatellite data (Bertola *et al.* submitted; Bertola *et al.* submitted). The obtained SNPs are analyzed in a phylogeographic framework. Compared to previously published phylogeographic patterns, based on mtDNA and microsatellite data, this provides a more complete overview of the complexity underlying intraspecific genetic diversity in the lion. Finally, a selection of the discovered SNPs can be used for a wider study on more sampling locations, potentially contributing to future studies on lion genetics.

Materials and Methods

Blood or tissue samples of ten lions, representing the main phylogeographic groups (Figure 1), were collected and stored in buffer solution (0.15 M NaCl, 0.05 M Tris-HCl, 0.001 M EDTA, pH = 7.5) at -20 °C. All included individuals were either free-ranging lions or captive lions with proper documentation of their breeding history. A sample from a leopard (*Panthera pardus orientalis*, captive) was included as an outgroup. All samples were collected in full compliance with specific legally required permits (CITES and permits related to national legislation in the countries of origin).

DNA was extracted using the DNeasy Blood & Tissue kit (Qiagen) following the manufacturer's protocol. The DNA was sequenced on 3 lanes of an Illumina HiSeq2000 to 99 bp paired end reads with 200-400 bp insert size (Leiden Genome Technology Center, Leiden, The Netherlands). In the first run, two individuals (Benin and Kenya) were tagged and pooled with leopard DNA as the outgroup (ratios 1:1:2 for Benin, Kenya and Leopard respectively). In the two following runs four individuals (Cameroon+Somalia+RSA+India and DRC+Zambia1+Zambia2+Namibia) were tagged and equimolarily pooled (Supplemental Table S1). Resulting reads were identified based on the unique adapter sequences.

The sequencing run containing Benin, Kenya and Leopard was repeated, since the first run produced read pairs with a severe drop in quality in the second read (Supplemental Figure S1). We hard-clipped these reads after the first 30 bp and added these data to the reads derived from a second run of the same samples. Quality control was performed using the FastQC tool (Andrews 2010) on the raw reads, and after removing adapter sequences with cutadapt (Martin 2011) and quality trimming with Sickle (Joshi & Fass 2011).

Samples Benin and RSA showed bimodal distributions of GC content per read and high average GC content compared to the other samples (55% and 45%, respectively, versus ~40%), indicating contamination with bacterial DNA. A nucleotide Blast search (Altschul *et al.* 1990) was done on a random selection of 10,000 reads per sample. Bacterial genomes of the highest hits were downloaded from GenBank (Supplemental Table S2) and reads for samples Benin and RSA were aligned against these using BWA (Li & Durbin 2009). Only unaligned reads were retained. In a second filtering step, only reads aligning to the reference genome of an Amur tiger (*Panthera tigris altaica*) (Cho *et al.* 2013) were included for downstream analyses. Re-analysis of the GC content distribution for these samples showed that these filtering steps eliminated the second peak (Supplemental Figure S2).

A reference genome was created by concatenating an Amur tiger assembly (Cho *et al.* 2013) and supplementing this with a lion mtDNA genome (30. Cameroon; Bertola *et al.*, submitted). Reads of lions and Leopard were aligned to this reference using BWA (Li & Durbin 2009).

Single Nucleotide Variant (SNV) calling was performed using SAMtools mpileup (Li *et al.* 2009) with default settings on Leopard (outgroup) separately and all lion samples jointly. SNV calling was executed excluding samples Benin and RSA, because of the influence of these samples on the available coverage per sample. We filtered calls based on their quality (phred score ≥20) and per-sample read depth (≥6 for Leopard, ≥3 for all lion samples). Sample alleles at variant sites were derived from Leopard calls and lion calls on non-contaminated samples (i.e. excluding Benin and RSA). This file was enriched with data for Benin and RSA from the joint calling including all lions for positions where enough coverage for all samples was available. All other positions were filled with ambiguous nucleotides (N). This procedure was repeated using only sites that were variant within the lion samples (i.e. excluding outgroups). Calling Y chromosomal SNVs in the eight male samples was done as described above, only on scaffolds supposedly located on the Y chromosome, identified by aligning all known Y chromosomal regions in cat (*Felis catus*) to the genomic data from Cho *et al.* (2013) (Supplemental Table S3). We configured SAMtools to assume a haploid genome and all positions with a heterozygote calling (22 out of 164) were discarded. The resulting sample alleles were serialized to FASTA format and served as input for the phylogenetic analysis. The complete pipeline used in this project and additional information is available at: https://git.lumc.nl/lgtc-bioinformatics/bertola-lion

Phylogenetic analyses were performed using MrBayes v.3.1.2 (Huelsenbeck & Ronquist 2001; Ronquist *et al.* 2012) and Garli (Zwickl 2006), using parameters determined by MrModeltest2 (v.2.3) (Nylander 2004). Branches receiving >0.95 PP in Bayesian analysis (MrBayes) and/or 70 bootstrap support in Maximum Likelihood (ML) analysis (Garli) were considered to be significantly supported. In addition a Principle Component Analysis (PCA) was executed in Genalex (Peakall & Smouse 2012) and R version 3.1.0, using prcomp. Isolation by Distance (IBD) analyses were performed in Genalex using 999 permutations (Peakall & Smouse 2012), excluding the contaminated samples Benin and RSA due to difficulties in estimating genetic distance with a high frequency of ambiguous nucleotides. Levels of differentiation (Fst) were calculated using Arlequin using 1023 permutations (Excoffier *et al.* 2005). The level of heterozygosity was assessed for each lion, taking into account the numbers of scored (non-ambiguous) nucleotides. Further, identified SNPs were attributed to a chromosome, following the genomic architecture in the tiger (Cho *et al.* 2013) and Bayesian analyses and PCA were repeated for individual chromosomes. In addition, mtDNA data were subjected to Bayesian and ML analysis, and PCA by using mitochondrial genomes as identified by Bertola *et al.* (submitted).

Results

The sequencing runs yielded a total of 628,716,470 reads and after quality control a total of 593,632,293 reads (94.4%) were retained for subsequent alignment (Supplemental Table S1). Filtering of variable positions between ten lions, one leopard and one tiger, yielded 44,627 variable positions, of which 18,457 positions were variable within the lion. Assuming identical chromosomal architecture in the lion as in the tiger, we find a strong relationship between discovered SNPs in this study and estimated chromosome sizes in the tiger (Cho *et al.* 2013) (Supplemental Table S4). On the Y chromosome 142 SNPs were identified compared to the outgroup species. Coverage plots for all individuals and all scaffolds illustrate the Y-chromosomal origin, since hardly any coverage is found for the females included (Supplemental Figure S3). Since only 1 Y chromosomal position is variable within the lion, this alignment was not further subjected to phylogenetic analyses. Mitochondrial genomes, consisting of 16,756 bp, excluding repetitive regions RS-2 and RS-3 (Jae-Heup *et al.* 2001), were added to the dataset. On the mtDNA 2,317 SNPs were identified, with 742 variable positions within the lion.

Phylogenetic analyses, based on all lion-specific SNPs, showed a hierarchical pattern in which the populations from the northern part of the lion range represent the most basal branches (Figure 2). Exclusion of the contaminated samples, Benin and RSA, which contained high numbers of missing values, did not influence the topology or support of the tree. Similarly, the exclusion of intermediate populations, i.e. DRC, Somalia and Kenya, did not change the overall topology of the tree. As was previously shown with mtDNA markers, the Asiatic subspecies shows a close genetic relationship to lions from West and Central Africa and does not have an outgroup position. The phylogenetic tree based on the mitochondrial genomes shows a basal dichotomy, although the branch containing the southern populations is not well supported. Phylogenetic trees and PCA from individual chromosomes show largely congruent patterns (Supplemental Figure S4).

IBD analyses showed a strongly significant correlation between genetic and geographic distance, both including and excluding the Asiatic subspecies (Supplemental Table 5). Since tree topology indicates a more gradual differentiation, in contrast to the basal dichotomy observed in the mtDNA, population differentiation was calculated, regarding the geographically intermediate populations Somalia and Kenya as either 1) North, 2) South or 3) Intermediate. Pairwise Fst values were significant ($P<0.05$) in all cases, except when Somalia and Kenya were included as Intermediate, in which case only North and South populations showed significant differentiation from each other.

Individual levels of heterozygosity were assessed and compared to previously published data from Bertola *et al.* (submitted) and (Dubach *et al.* 2013) (Supplemental Table 6). Ranking of these levels between SNP data and microsatellite data finds strong congruence, although contaminated samples Benin and RSA had to be excluded due to the low coverage, which may bias the number of heterozygote positions.

Discussion

This study shows how whole genome sequencing can be used for SNP discovery in a non-model species, in this case the lion. The results are used in a phylogenetic framework to infer evolutionary histories of the lion and compare these results to previously published scenarios. Since this approach mainly served the identification of variable positions in the lion, the number of included samples for a phylogenetic analysis is restricted. However, the identified SNPs can be used as a source for the generation of a SNP panel, based on which a larger number of individuals can be genotyped.

Previously published mtDNA datasets of the lion showed a strongly supported basal dichotomy, clustering all populations from the northern part of the range, including the Asiatic subspecies, and all populations from the southern part of the range. Although the branch with southern populations did not receive significant support when only ten individuals were included (Figure 2), we interpret the tree as having a basal dichotomy, as was previously shown by Barnett *et al.* (2014) and Bertola *et al.* (submitted). This basal dichotomy is less pronounced in the SNP data, notably due to the structure in the northern part of the range. However, the Asiatic subspecies is nested in the African lion tree, close to lions from West and Central Africa, further undermining the validity of its distinct subspecies status. The hierarchical pattern observed in the SNP data may largely be attributed to continent-wide gene flow, explaining the more gradual pattern of population differentiation. The consecution in which the individuals branch off, support this explanation.

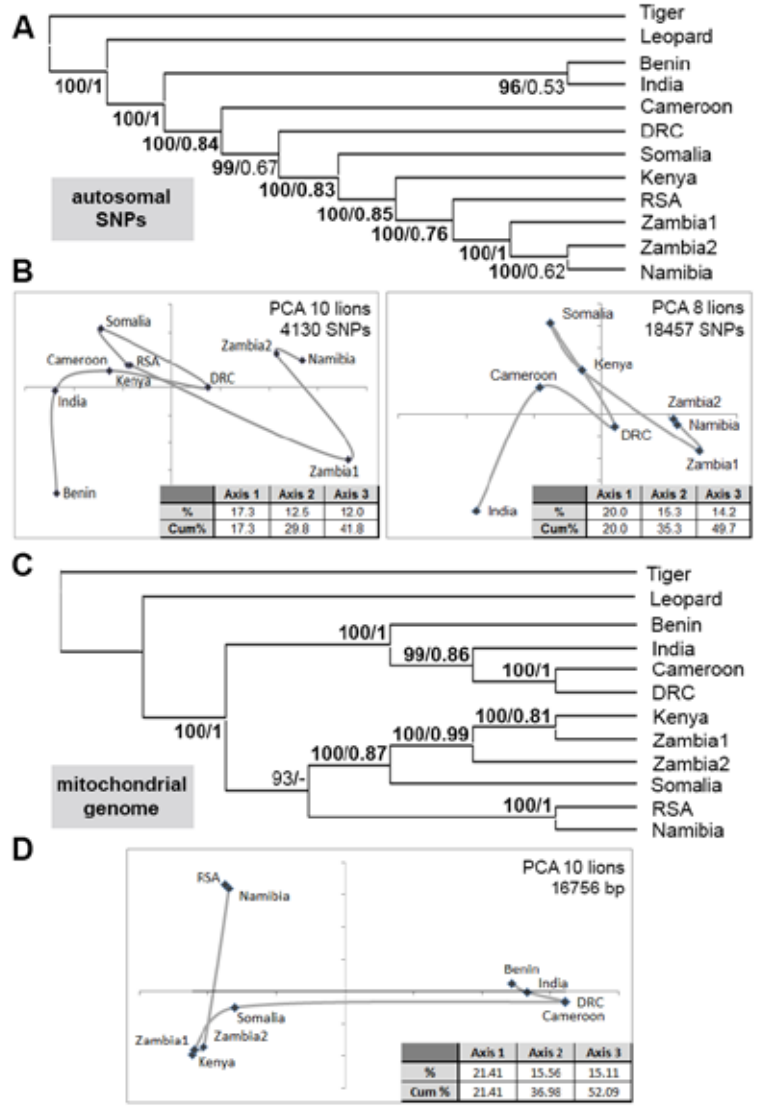


Figure 2. Bayesian analysis and PCA of SNPs in lion, leopard and tiger. A: Bayesian and ML analysis of 18,457 SNPs in ten lions, with posterior probability/bootstrap values indicated at the nodes. B: PCA of all variable positions in the lion, including and excluding the contaminated samples Benin and RSA. The line connects the populations in the same order as the topology of the tree. C: Bayesian and ML analysis of complete mitochondrial genomes of 10 lions, with posterior probability/bootstrap values indicated at the nodes. D: PCA based on the complete mitochondrial genome of ten lions. The line connects the populations in the same order as the topology of the tree.

Since dispersal in lions is biased to the male sex (Pusey *et al.* 1987; Spong & Creel 2001), this may explain why we see a more discrete phylogenetic pattern in the mtDNA. Major barriers for gene flow seem to be restricted to the (recent) population gap in North Africa/Middle East and the Central African rain forest. Although the Rift valley has frequently been mentioned as a barrier for gene flow in the lion (Dubach *et al.* 2005; Barnett *et al.* 2006a; b; Bertola *et al.* 2011a), and gene flow may be reduced, admixture between haplogroups indicates that the Rift valley is not a complete barrier for

lion dispersal (Bertola et al, submitted). In historic times, additional barriers may have existed as a result of expanding rain forest or desert (Bertola *et al.* submitted). The restriction of suitable lion habitat to a small number of refugia may have contributed to the development of discrete genetic lineages. The pattern found in mtDNA data of the lion is congruent with that of other species (Hewitt 2004; Lorenzen *et al.* 2012; Bertola *et al.* submitted) and predicted refugial areas based on climate models (Levinsky *et al.* 2013). Faster coalescence times of mtDNA may have led to reciprocally monophyletic mtDNA clades in the lion, while isolation in refugia may not have lasted long enough for coalescence in autosomal markers, due to the cyclic character of the African climate (Bertola et al, submitted).

Autosomal SNPs and microsatellite data are expected to produce largely congruent patterns because of a similar mode of inheritance and coalescence times. Due to the hierarchical nature of the SNP tree it is difficult to interpret which groups can be considered to be discrete. SNP data may represent a more ancient pattern, in which historic gene flow is strongly represented, while phylogeographic patterns based on microsatellite data may, as a result of their high mutation rate, represent relatively recent evolutionary history, as is the case for fast coalescent markers, like mtDNA. This may explain why distinct clusters are relatively easily retrieved from microsatellite data, but not from SNP data.

Based on microsatellites population Zambia1 was indicated as an admixture zone. IBD analysis from the SNP data seem to confirm this: notably after exclusion of India, Zambia1 forms a relatively distinct cloud, representing low genetic distance, compared to the other pairwise comparisons. We do not find indications for a suture zone between mtDNA haplogroups in this region, indicating the admixture may be the result of male-biased gene flow. An admixture pattern of haplogroups is found in Ethiopia, where the presence of a suture zone is further supported by microsatellite data (Bertola *et al.* submitted). Based on current sampling locations in DRC, Kenya and Somalia, their position in the PCA plots and the formation of a loop connecting these sampling localities also suggests admixture. SNP data from more sampling localities in this region may be able to further support this. Finally, the position of RSA in the PCA plot may be the result of human-mediated admixture in RSA, visible as a mosaic pattern of haplogroups in the Kruger/Limpopo area. This individual contains a South West haplotype, but is likely to be admixed with East/Southern African lions, which explains the close position to Kenya.

Ranking observed heterozygosity values, results in a congruent pattern between SNP and microsatellite data. This supports the notion that, even though a single individual per population has been sampled for the SNP discovery, the number of SNPs identified can give an indication of within-population diversity levels. SNP genotyping for more individuals from a single population could be executed to further strengthen this point.

A genome wide SNP panel, based on ten lions from the main phylogeographic groups, shows a gradual degree of relatedness of lions following a north-south axis, and a nested position of the Asiatic lion within the African subspecies. This suggests that the current nomenclature, recognizing an African and an Asiatic subspecies, conflicts with the distribution of genetic diversity in the species, as was previously shown for mtDNA data only (Dubach *et al.* 2013; Barnett *et al.* 2014; Bertola *et al.* submitted). Although the phylogeographic pattern based on genome-wide autosomal markers is more gradual, without recognizing reciprocally monophyletic clades, suggestions regarding management of lion populations postulated by Barnett *et al.* (2014) and Bertola *et al.* (submitted) still hold. Defining units for conservation management by looking for reciprocal monophyly in autosomal data may be overly restrictive. Following current insights, combining mtDNA, microsatellite and genome-wide SNP data, we confirm six ESUs as previously suggested based on reciprocally monophyletic haplogroups: 1) West Africa, 2) Central Africa, 3) India, 4) North East Africa, 5) East/Southern Africa and 6) South West Africa. Finally, due to the nested position of the Asiatic subspecies, we support a taxonomic revision, distinguishing an northern subspecies, including the Asiatic lion, and a southern subspecies in the lion. Based on the discovered SNPs from this paper a SNP panel has been designed, also including mitochondrial SNPs, which can be used for fast and cost-effective genotyping of large numbers of individuals. This method may also be applied for the establishment of breeding programmes for captive stocks or in a forensics framework to trace source populations of illegal lion products. Analysing more free-ranging lion populations will further improve the understanding of their levels of diverstiy, genetic relationships and evolutionary history.

Acknowledgements

Samples were kindly provided by E. Sogbohossou (Benin), P. Tumenta, S.Adam, R. Buij and B. Croes (Cameroon), ICCN, Garamba NP (DRC), Safaripark Beekse Bergen (Hilvarenbeek, The Netherlands) (Somalia), B. Patterson (Kenya), O. Aschenborn (Namibia), Ouwehands Dierenpark (Rhenen, The Netherlands) (RSA), C.A.Driscoll (India) and Planckendael (Muizen, Belgium) (Leopard). We further thank N. Schidlo, H. Buermans, Y. Ariyurek, and S. Greve-Onderwater for assisting in processing of the samples and Cho et al. for assistance with the reference data. The investigations were supported by the Division for Earth and Life Sciences (ALW) with financial aid from the Netherlands Organization for Scientific Research (NWO) (project no. 820.01.002).

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic Local Alignment Search Tool 2Department of Computer Science. *Journal of Molecular Biology*, **215**, 403–410.

Andrews S (2010) FastQC - A Quality Control tool for High Throughput Sequence Data.

Antunes A, Troyer JL, Roelke ME *et al.* (2008) The evolutionary dynamics of the lion Panthera leo revealed by host and viral population genomics. *PLoS genetics*, **4**, e1000251.

Barnett R, Yamaguchi N, Barnes I, Cooper A (2006a) Lost populations and preserving genetic diversity in the lion Panthera leo: Implications for its ex situ conservation. *Conservation Genetics*, **7**, 507–514.

Barnett R, Yamaguchi N, Barnes I, Cooper A (2006b) The origin, current diversity and future conservation of the modern lion (Panthera leo). *Proceedings of the Royal Society B: Biological Sciences*, **273**, 2119–2125.

Barnett R, Yamaguchi N, Shapiro B *et al.* (2014) Revealing the maternal demographic history of Panthera leo using ancient DNA and a spatially explicit genealogical analysis. *BMC Evolutionary Biology*, **14**, 70.

Bertola LD, van Hooft WF, Vrieling K *et al.* (2011) Genetic diversity, evolutionary history and implications for conservation of the lion (Panthera leo) in West and Central Africa. *Journal of Biogeography*, **38**, 1356–1367.

Cho YS, Hu L, Hou H *et al.* (2013) The tiger genome and comparative analysis with lion and snow leopard genomes. *Nature communications*, **4**.

Dubach JM, Briggs MB, White PA, Ament BA, Patterson BD (2013) Genetic perspectives on "Lion Conservation Units" in Eastern and Southern Africa. *Conservation Genetics*, **1942**.

Dubach J, Patterson BD, Briggs MB *et al.* (2005) Molecular genetic variation across the southern and eastern geographic ranges of the African lion, Panthera leo. *Conservation Genetics*, **6**, 15–24.

Edwards SV (2009) Is a new and general theory of molecular systematics emerging? *Evolution; international journal of organic evolution*, **63**, 1–19.

Edwards S, Bensch S (2009) Looking forwards or looking backwards in avian phylogeography? A comment on Zink and Barrowclough 2008. *Molecular ecology*, **18**, 2930–3; discussion 2934–6.

Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evolutionary bioinformatics online*, **1**, 47–50.

Hewitt GM (2004) The structure of biodiversity - insights from molecular phylogeography. *Frontiers in zoology*, **1**, 4.

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics (Oxford, England)*, **17**, 754–5.

Jae-Heup K, Eizirik E, O'Brien SJ, Johnson WE (2001) Structure and patterns of sequence variation in the mitochondrial DNA control region of the great cats. *Mitochondrion*, **1**, 279–292.

Joshi NA, Fass JN (2011) Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files.

Knowles LL (2009) Statistical Phylogeography. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 593–612.

Levinsky I, Araújo MB, Nogués-Bravo D *et al.* (2013) Climate envelope models suggest spatio-temporal co-occurrence of refugia of African birds and mammals. *Global Ecology and Biogeography*, **22**, 351–363.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, **25**, 1754–60.

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, **25**, 2078–9.

Lorenzen ED, Heller R, Siegismund HR (2012) Comparative phylogeography of African savannah ungulates. *Molecular ecology*, **21**, 3656–70.

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EBMnet.journal*, **17**, 10–12.

Miller SM, Bissett C, Burger A *et al.* (2013) Management of reintroduced lions in small , fenced reserves in South Africa : an assessment and guidelines. *South African Journal of Wildlife Research*, **43**, 138–154.

Moritz C (1994) Defining "Evolutionarily Significant Units" for conservation. *Trends in ecology & evolution (Personal edition)*, **9**, 373–375.

Nylander JAA (2004) MrModeltest v2.3.

Peakall R, Smouse PE (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research--an update. *Bioinformatics (Oxford, England)*, **28**, 2537–9.

Pusey AE, Packer C, Erhoff-Mulder MB (1987) The Evolution os Sex-biased Dispersal in Lions. *Behaviour*, **101**, 275–310.

Ronquist F, Teslenko M, van der Mark P *et al.* (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology*, **61**, 539–42.

Spong GF, Creel S (2001) Deriving dispersal distances from genetic data. *Proceedings. Biological sciences / The Royal Society*, **268**, 2571–2574.

Zink RM, Barrowclough GF (2008) Mitochondrial DNA under siege in avian phylogeography. *Molecular ecology*, **17**, 2107–2121.

Zwickl DJ (2006) Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. The University of Texas at Austin.

Data accessibility

The complete pipeline used in this project and additional information is available at: https://git.lumc.nl/lgtc-bioinformatics/bertola-lion

Supporting Information

Supporting information which is not included here may be found in the online version of this article and is available upon request.

Supplemental Table S1. Lion and leopard samples included in this study and results whole genome sequencing.

Run	Sample	Corresponding to Bertola et al. (submitted)	Origin	Sex	Read number	Read number after filtering**	Read number after filtering + QC	% left after QC	GC%
1	**Leopard**	196.Leopard	Amur (captive[1])	Male	172,038,427 (+ 153,603,680 partially failed*)	-	-	94	40
	Benin	9.Benin	Benin - Pendjari NP	Female	18,698,767 (+ 16,318,397 partially failed*)	11,846,142 (+ 9,184,453 partially failed)	11,665,188 (+ 8,834,199 partially failed)	94	55 (bimodal)
	Kenya	89.Kenya	Kenya - Tsavo East NP	Male	32,033,356 (+ 28,473,401 partially failed*)	-	-	94	40
2	**India**	174.India	India - Gir forest[2]	Male	32,453,394	-	-	98	40
	Cameroon	21.Cameroon	Cameroon - Waza NP	Male	37,995,410	-	-	98	40
	Somalia	71.Somalia	Somalia (captive[3])	Male	20,683,546	-	-	94	40
	RSA	162.RSA	RSA (captive[4])	Female	29,190,629	-	-	98	45 (bimodal)
3	**DRC**	42.DRC	DRC - Garamba NP	Male	29,552,205	-	-	97	42
	Zambia1	95.Zambia	Zambia - Luangwa Valley	Male	22,700,633	-	-	98	38
	Zambia2	96.Zambia	Zambia - Mulobezi town	Male	25,796,481	-	-	97	42
	Namibia	131.Namibia	Namibia - Etosha NP	Male	25,496,541	-	-	98	39
					160,870,954 (+ 138,506,569 partially failed*)	-	-	94	40
					17,557,523 (+ 13,741,759 partially failed*)	11,846,142 (+ 9,184,453 partially failed)	11,665,188 (+ 8,834,199 partially failed)	94	55 (bimodal)
					30,217,372 (+ 26,035,900 partially failed*)	-	-	94	40
					31,687,608	-	-	98	40
					37,257,634	-	-	98	40
					19,527,670	-	-	94	40
					28,468,409	28,262,272	28,074,951	98	45 (bimodal)
					28,718,622	-	-	97	42
					22,188,079	-	-	98	38
					25,141,802	-	-	97	42
					24,905,745	-	-	98	39

* "partially failed" indicates the number of reads from a previous run, which were trimmed to 30 bp in the second read only due to a severe drop in quality.

** Filtering refers to filtering of reads from bacterial origin in two contaminated samples, Benin and RSA.

1 Planckendael: Planckendael, Muizen, Belgium

2 Captive born, founders both wild; Sakkarbaug Zoo: Sakkarbaug Zoological Garden, Junagadh, Gujarat, India

3 Safaripark Beekse Bergen, Hilvarenbeek, The Netherlands

4 Ouwehands dierenpark, Rhenen, The Netherlands

Supplemental Table S2. Genbank entries to filter bacterial reads in contaminated samples Benin and RSA.

Genbank Accession	Organism	Details
gi|386716467|ref|NC_017671.1|	Stenotrophomonas maltophilia D457	complete genome
gi|206558403|ref|NC_011000.1|	Burkholderia cenocepacia J2315	chromosome 1, complete sequence
gi|206561868|ref|NC_011001.1|	Burkholderia cenocepacia J2315	chromosome 2, complete sequence
gi|191639869|ref|NC_011002.1|	Burkholderia cenocepacia J2315	chromosome 3, complete sequence
gi|206479926|ref|NC_011003.1|	Burkholderia cenocepacia J2315	plasmid pBCJ2315, complete sequence

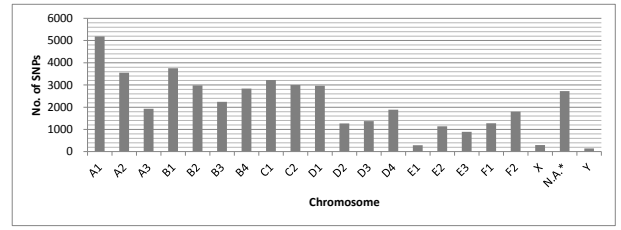
Supplemental Table S3. Scaffolds in the reference sequence (Cho *et al.*, 2013) identified as potentially from Y-chromosomal origin.

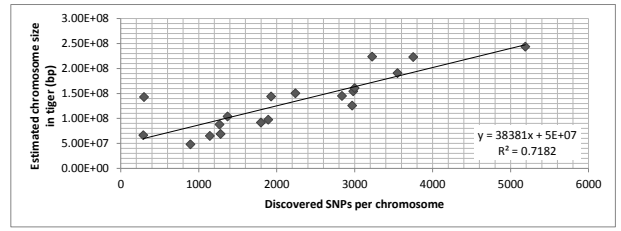
scaffold	Gene	gi	Score	E-value
scaffold725	SRY	77176790	4149	0
scaffold638	UBE1Y	84620608	549	e-153
scaffold363	CYorf15	84620610	172	8.00E-41
scaffold640	CUL4BY	84620611	696	0
scaffold1087	TETY2	84620617	975	0

Supplemental Table S4. Number of discovered SNPs per chromosome and estimated chromosome size in tiger (Cho *et al.*, 2013).

Chromosome	Discovered SNPs	Estimated chromosome size (bp) Amur tiger (Cho et al., 2013)
A1	5,188	243,492,181
A2	3,550	190,495,254
A3	1,929	144,011,757
B1	3,750	222,683,385
B2	2,982	154,295,958
B3	2,240	150,246,213
B4	2,837	144,888,701
C1	3,225	223,586,761
C2	3,000	160,670,131
D1	2,969	125,709,129
D2	1,267	87,703,667
D3	1,371	103,759,264
D4	1,891	97,290,273
E1	290	66,408,731
E2	1,143	64,743,307
E3	893	47,874,673
F1	1,280	68,695,903
F2	1,798	91,576,383
X	298	142,585,357
N.A.*	2,726	-
Y	142	-

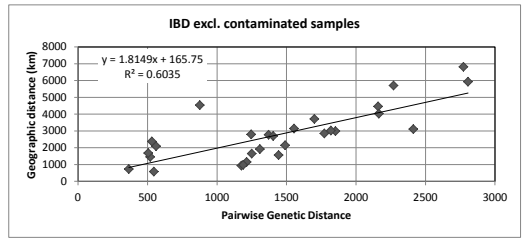
* SNPs which could not be assigned to any of the chromosomes

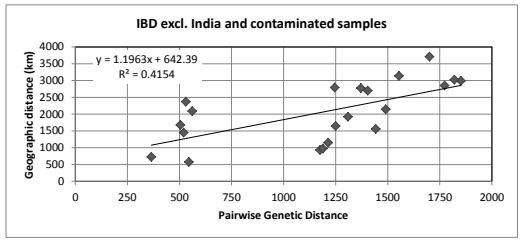




Supplemental Table S5. IBD analysis for 8 lion samples (excluding the contaminated samples Benin and RSA) and for 7 samples (excluding India and the contaminated samples).

Groups	Pair. comparis	Rxy
8 samples (excl. contaminated)	28	0.777 (P≤0.002)
7 samples (excl. India + contaminated)	21	0.645 (P≤0.001)

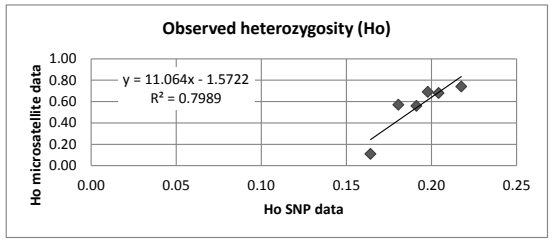


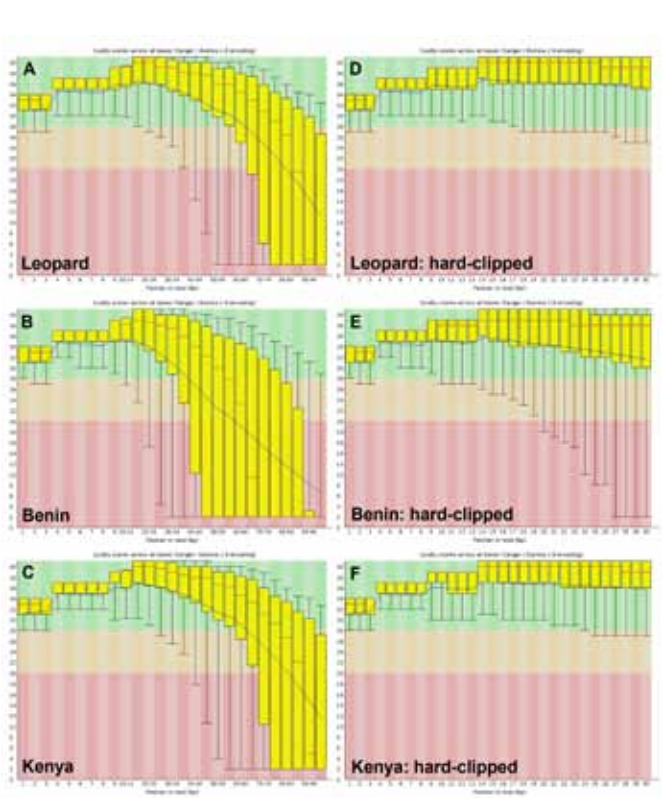


Supplemental Table S6. Observed heterozygosity for all lion samples and comparison with observed heterozygosity based on microsatellite data. Shading indicates the ranking from low heterozygosity (red) to high heterozygosity (green).

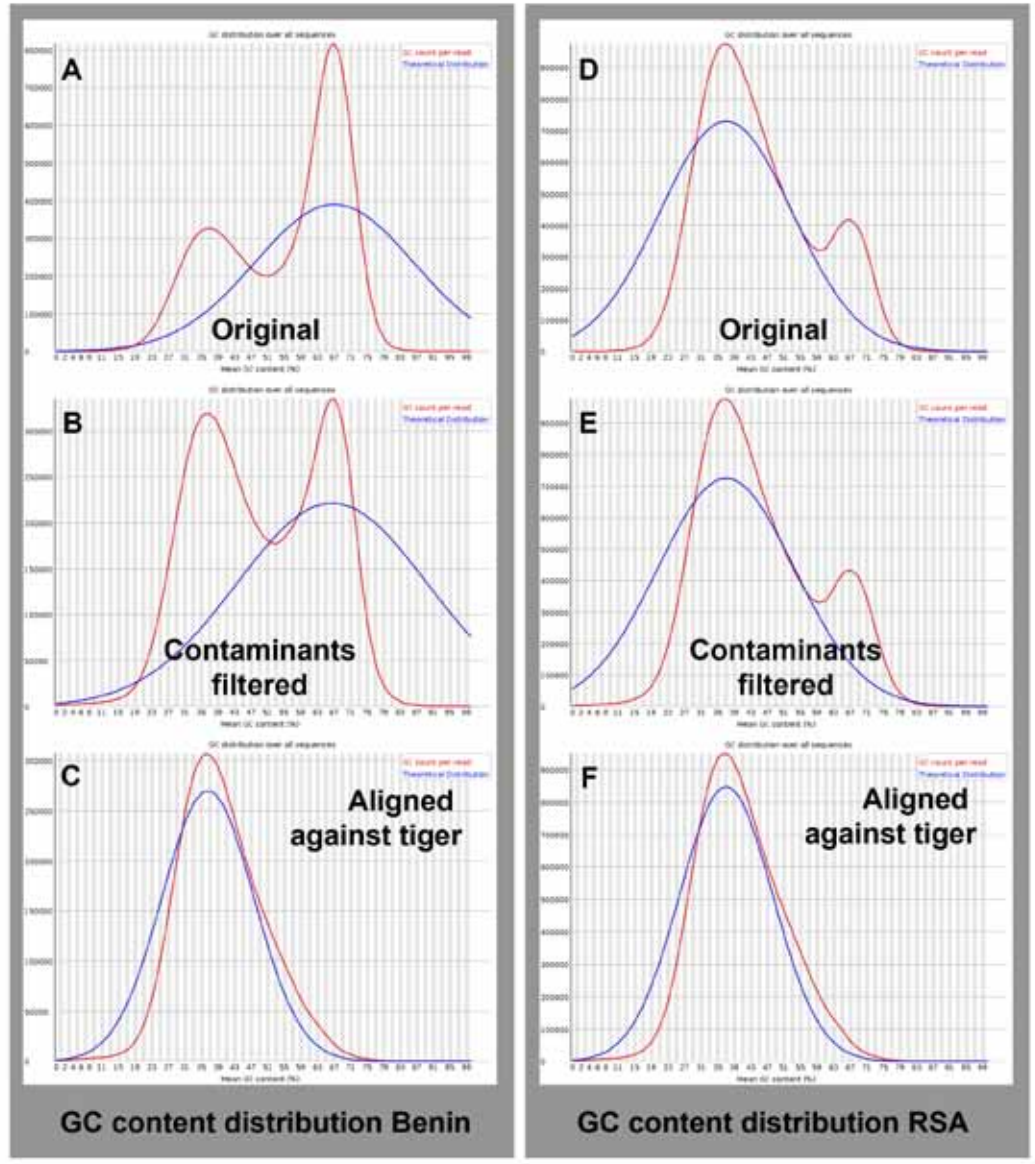
Sample	SNPs scored (non-ambiguous)	heterozygote positions	homozygote positions	Observed heterozygosity (Ho)	Ho based on microsatellite data	Source microsatellite data
Benin	8,106	2,157	5,949	0.27*	0.65	Bertola et al., submitted
India	44,627	7,326	37,301	0.16	0.11	Bertola et al., submitted
Cameroon	44,627	9,111	35,516	0.20	0.68	Bertola et al., submitted
DRC	44,627	9,707	34,920	0.22	0.74	Bertola et al., submitted
Somalia	44,627	8,092	36,535	0.18	-	-
Kenya	44,627	10,004	34,623	0.22	-	-
Zambia1	44,627	8,057	36,570	0.18	0.57	Bertola et al., submitted
Zambia2	44,627	8,828	35,799	0.20	0.69	Dubach et al., 2013
RSA	20,667	5,333	15,334	0.26*	0.69	Bertola et al., submitted
Namibia	44,627	8,532	36,095	0.19	0.56	Bertola et al., submitted

* Contaminated samples were excluded from the ranking due to low coverage

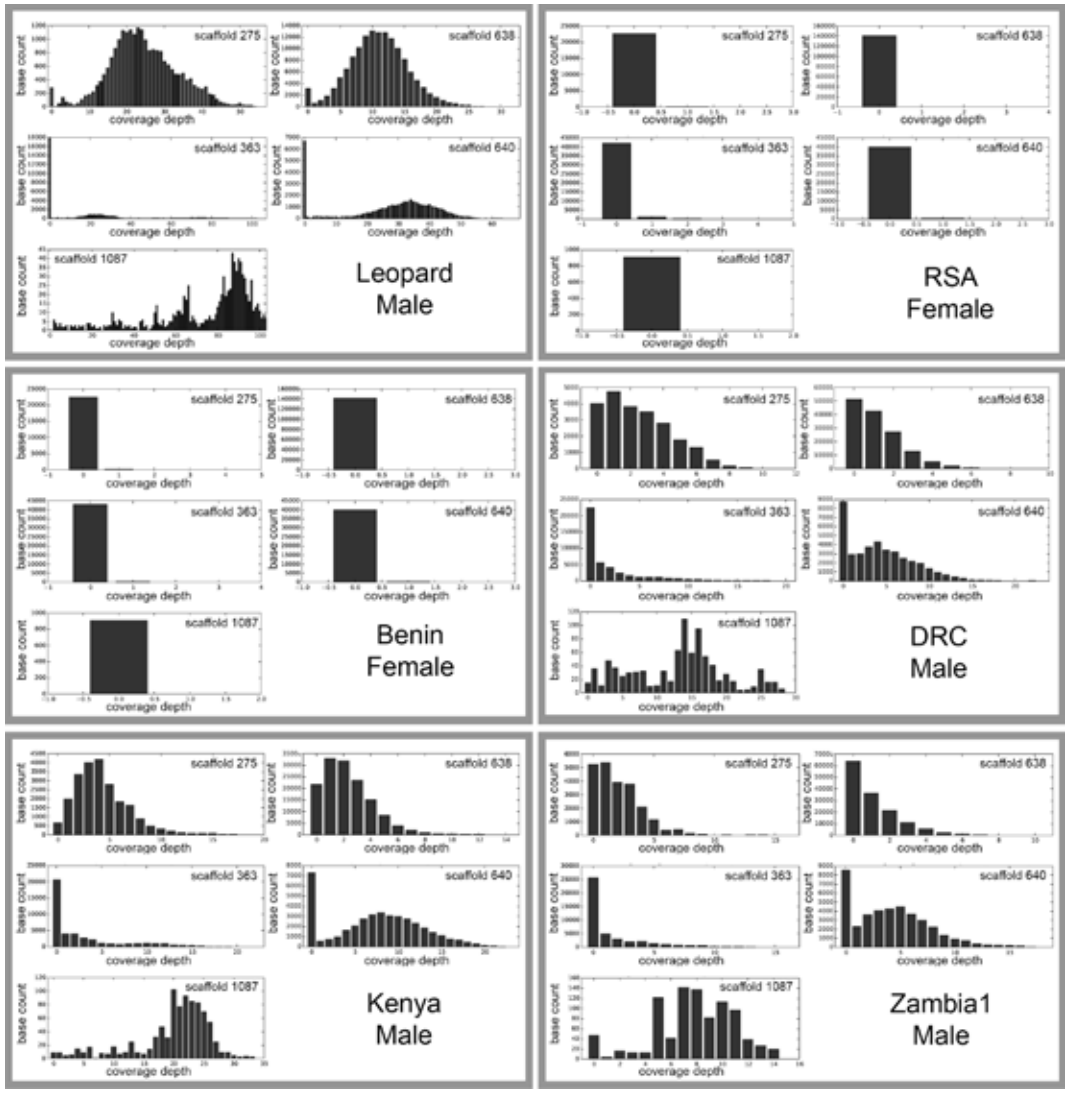


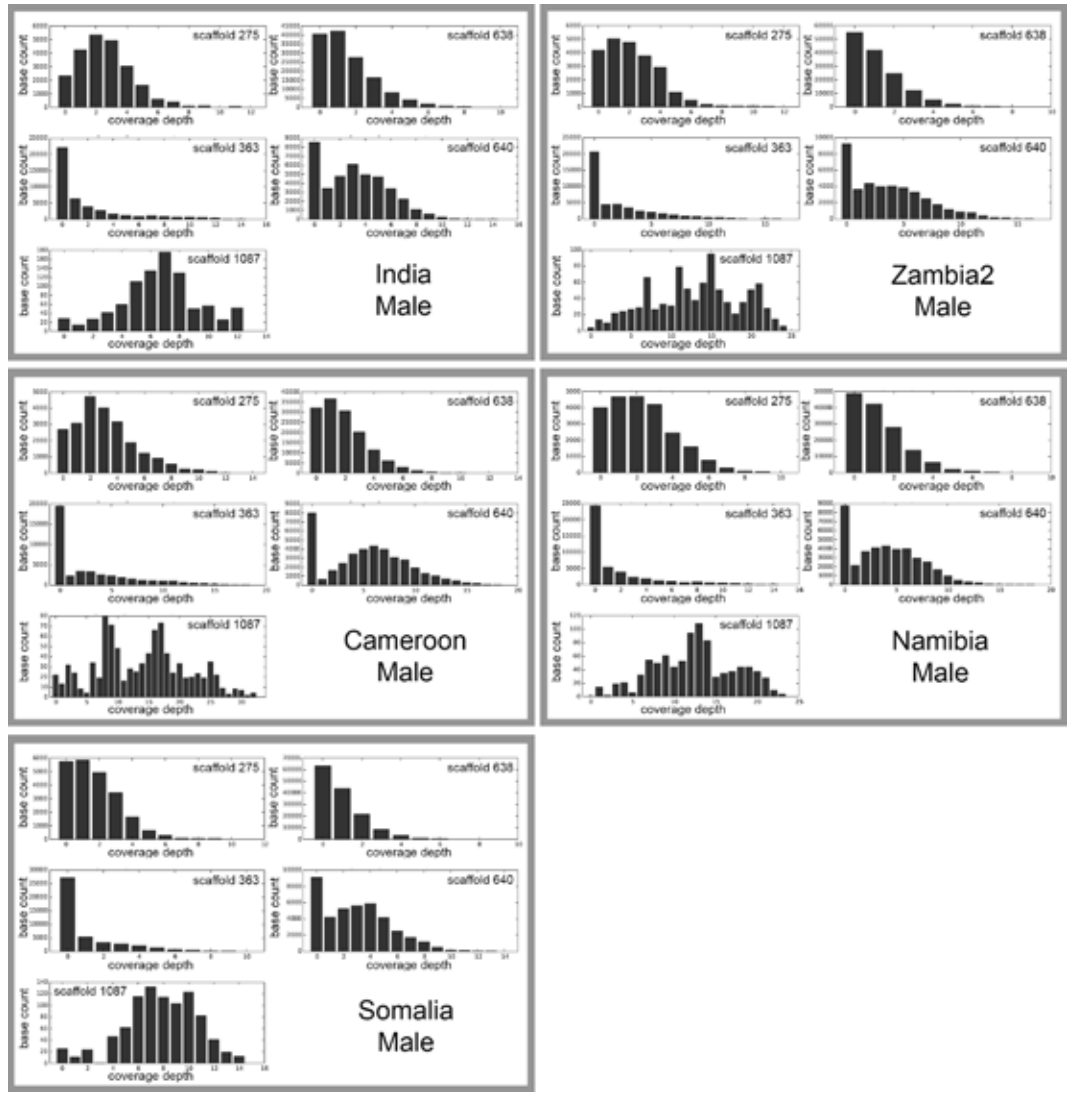


Supplemental Figure S1. Read quality derived from the first run, containing one leopard and two lion samples. Drop in quality scores for (A) Leopard, (B) Benin and (C) Kenya and quality after hard clipping of reads after 30 bp for (D) Leopard, (E) Benin and (F) Kenya.

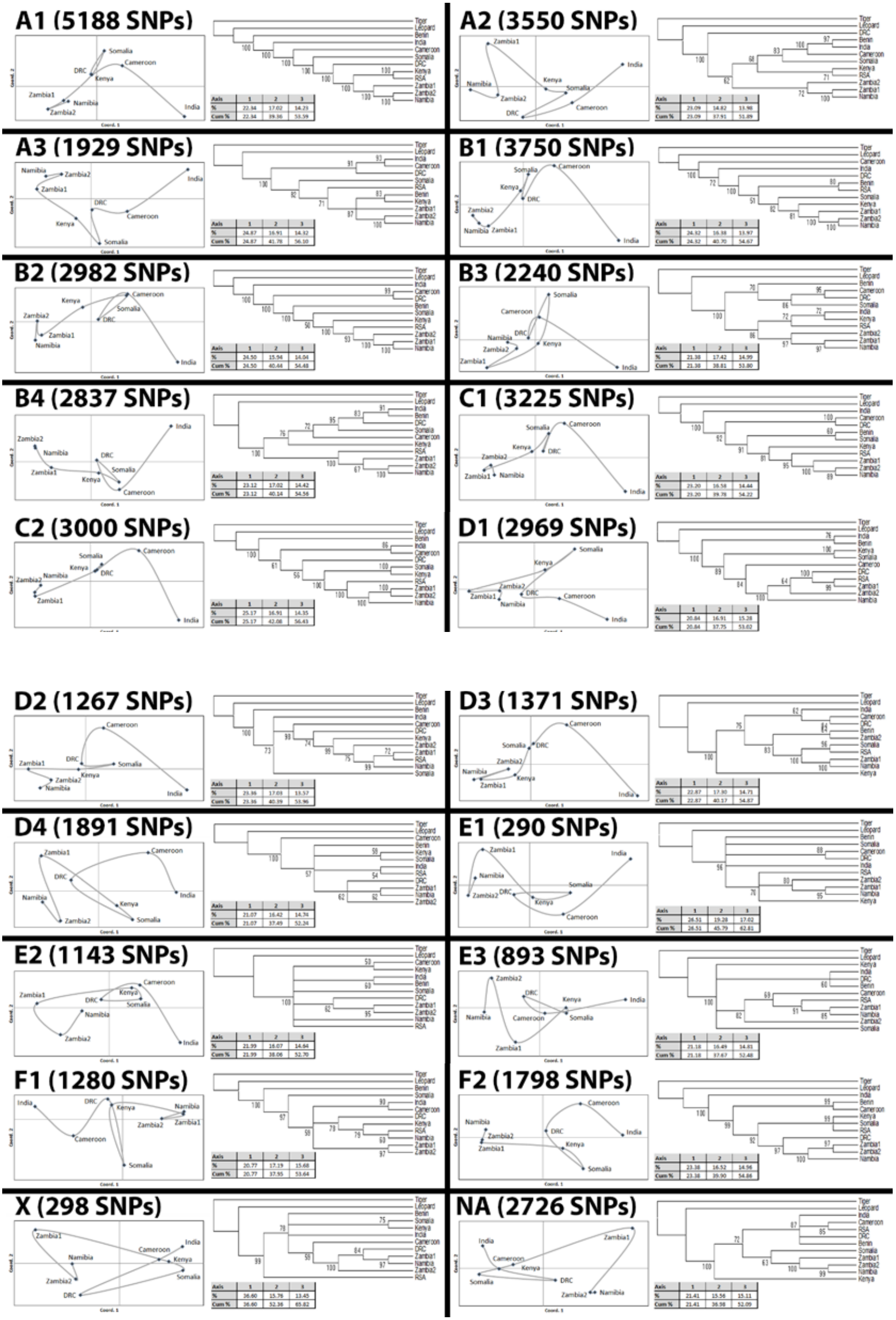


Supplemental Figure S2. GC content distribution for two lion samples showing signs of bacterial contamination. GC content of raw reads of (A) Benin and (D) RSA, (B+E) reads filtered against main contaminants and (C+F) reads aligned again the reference genome of the tiger.





Supplemental Figure S3. Coverage plots for one leopard and ten lions on five scaffolds that had been identified as having an Y-chromosomal origin.



Supplemental Figure S4. Bayesian analyses and PCA for individual chromosomes in the lion.