

Supplementary Methods

De novo assembly and annotation of Asiatic lion (*Panthera leo persica*) genome

Siuli Mitra^{1,2}, Ara Sreenivas¹, Divya Tej Sowpati³, Amitha Sampat Kumar³, Gowri Awasthi^{4,5}, M. Milner Kumar^{4,6}, Wajeeda Tabusum¹, **Ajay Gaur^{1*}**

1. Laboratory for Conservation of Endangered Species (LaCONES), CSIR-Centre for Cellular and Molecular Biology, Attapur, Hyderabad, India 500048
2. Present address Translational Health Science and Technology Institute (THSTI), Faridabad, India 121001
3. Genome Organisation and Nuclear Architecture Lab, CSIR-Centre for Cellular and Molecular Biology, Uppal Road, Hyderabad, India 500007
4. Past address: SciGenome Labs, Kochi, India 682037
5. Present address: Yashraj Biotechnology Ltd. Navi Mumbai, India 400705
6. Present address: Indian Institute of Technology, Madras, Chennai, India 600036

* **Corresponding Author (agaur@ccmb.res.in)**

Preparation of sample for sequencing:

The blood sample of an Asiatic lion individual named Atul was obtained from Nehru Zoological Park, Hyderabad (vide permission letter from Chief Wildlife Warden and Special Principal Chief Conservator of Forests-Dev., Telangana, letter Rc. No. 43505/2013/WL-1, dated 12/11/2014, **(Supplementary Table S1)**). DNA was isolated using the lysis method of NucleoSpin® Blood L (Machery-Nagel, GmbH & Co., Germany) following the manufacturer's recommendations. The quality and quantity of the sample was checked using NanoDrop ND 1000 (ThermoFisher Scientific, USA), Qubit (ThermoFisher Scientific, USA) and gel electrophoresis **(Supplementary Table S2 and Figure L1)** to make sure that the DNA obtained is sufficient and of superior quality for library preparation.

Asiatic lion genome sequencing:

Libraries of different insert sizes were constructed at SciGenom Labs, Cochin, India. Three short insert paired-end libraries (150bp, 500bp and 800bp) and three long insert mate-pair libraries (4-6Kb, 8-10Kb, and 1-20Kb) were sequenced using HiSeq 2500 (Illumina Inc., USA) **(Supplementary Table S3)**. A total of 231 Gb (Gigabase) data was generated as a result.

Raw read filtering for *de novo* assembly:

The post-sequencing analysis was earmarked with filtering steps for data quality control and comprised removal of low quality reads, trimming of adapter sequences and sequencing error correction to avoid a low-quality genome assembly. After these steps, the six libraries carrying varied insert sizes remained with that of 152 Gb data **(Supplementary Table S4)**.

***De novo* assembly of the Asiatic lion genome:**

The *de novo* assembly of the filtered reads was carried out using the De Bruijn graph-based approach in the SOAPdenovo software⁴² (<http://soap.genomics.org.cn/soapdenovo.html>). The contigs were further organized into scaffolds. Scaffolding was done using SOAP scaffold and SSPACE⁴⁴ with the help of the domestic cat genome owing to its phylogenetic proximity to the Asiatic lion and being the only felid with a chromosome level assembly. A total of 2.46 GB of scaffolds were assembled.

The genome size of Asiatic lion was estimated by using varied K-mer sizes of 17bp, 21bp and 23bp (**Supplementary Tables S5 and S6**). The estimation was carried out using BMAP (<https://sourceforge.net/projects/bbmap/>).

Asiatic Lion blood Transcriptome:

The quality of the Asiatic lion genome assembly was evaluated by sequencing the transcriptome of five Asiatic lion individuals: Ajay, Atul, Rita, Sonia and Viswas using HiSeq2000 (Illumina Inc. USA). The Trinity algorithm⁶⁵ (<http://trinityrnaseq.sourceforge.net/>) was applied to assemble the Asiatic lion transcriptome as the latter is specifically used for *de novo* assembly of transcripts (**Supplementary Table S7**). Trinity was chosen as it recovers full-length transcripts with efficiency that is equal across varied expression levels. It depends on the De Bruijn graphs to reconstruct the transcriptome by recovering alternate spliced isoform and recently duplicated genes.

Following that, Blat was used to align the assembled transcripts with the Asiatic lion assembly. The proportion of assembled transcripts covered in the Asiatic lion scaffolds was about 81% **(Supplementary Table S8)**.

Another step taken to examine the quality of the assembly, sequences of domestic cat EST were recovered from (<http://hgdownload.cse.ucsc.edu/goldenPath/felCat4/bigZips/est.fa.gz>) to be mapped to the Asiatic lion genome sequence. About 86% of the domestic cat ESTs was covered by the lion assembly while the mapped percent was quite high at 96.63% **(Supplementary Table S9)**. Additionally, Asiatic lion draft genome assembly analyzed by core eukaryotic genes using CEGMA⁴³ approach revealed that >86.8% of conserved genes in the assembly **(Supplementary Table S10)** indicating that good coverage and completeness with unique gene sequences in Asiatic lion genome assembly and it is same as that observed in the African lion assembly. The GC content of the Asiatic lion genome was determined to be 39.63% close to 41.66% obtained for a domestic cat and 41.4% in Amur tiger suggesting less bias due to assembly quality **(Supplementary Table S11)**.

Draft Chromosomes of Asiatic lion:

The draft chromosomes of Asiatic lion were constructed by mapping it to the reference genome of domestic cat (*Felis_catus* 8.0). The phylogenetic proximity of the two species is suggestive of the fact that the chromosomal location and ordering information of the lion scaffolds could be derived from the domestic cat genome **(Supplementary Table S12)**.

Homozygous variations (88% of SNVs and 11% of indels), found by mapping the Asiatic lion short reads to Asiatic lion scaffolds, were replaced with the nucleotides of the cleaned reads (**Supplementary Tables S13 and S14**). We aligned the short reads of Asiatic lion to the scaffolds using Bowtie2-2.2.5⁶⁶ with very sensitive option, and variations using the SAMtools-1.2⁶⁷. Further, when the Asiatic lion raw reads were mapped to the Asiatic lion scaffold 96.65% of the reads mapped to the lion assembly generated earlier. Also, 221,397 homozygous SNVs (73.85%) and 22,790 heterozygous SNVs (7.60%) were detected.

The corrected Asiatic lion scaffolds were mapped to the domestic cat genome (*Felis_catus-8.0*), using Burrows Wheeler Aligner, BWA⁶⁸, to determine the chromosomal location and order of the scaffolds. Two overlapping scaffolds without any mismatch were merged using megamerger of EMBOSS package⁶⁹ and non-overlapping scaffolds were joined using 100 'X' characters. In addition, the Asiatic lion short reads were aligned to the domestic cat genome using BWA with default options, and continuous mapped regions (consensus sequences) were extracted. A total of 41,532,416 homozygous SNVs were used to generate the consensus sequences (**Supplementary Table 15**). Finally, 122,195 of 134, 284 scaffolds (91% of the total scaffold length), were placed in the Asiatic lion chromosomes. (**Supplementary Tables S15 and S16**).

Assembly quality:

The N50 is the most appropriate parameter to adjudicate the contiguity of an assembly and is the weighted mean of assembled contigs. The lion genome was assembled and contigs were constructed using CLC Genomics Workbench 9.5.3 (CLC bio, Aarhus, Denmark, <https://www.qiagenbioinformatics.com/>), which utilizes the SIMD instructions to

parallelize and accelerate the assembly algorithms, making the program the fastest next generation sequencing assembler at present. For scaffolding of the contigs, SSPACE was used and using this software, a total of 2.46 Gb of scaffolds were assembled. The N50 of the scaffolds was 36,647bp, and the size of the longest scaffold was 383,982 bp (**Supplementary Table S6**).

Repeat annotation:

Repeated sequences are patterns of nucleic acids (DNA or RNA) that occur in multiple copies throughout the genome. They fall into three broad categories: (a) Terminal (Simple) repeats, (b) Tandem repeats and, (c) Interspersed repeats. Here, we have analyzed the microsatellite distribution in Asiatic lion using a comprehensive SSR mining tool, PERF¹⁸ and also Tandem Repeat Finder (TRF) version 4.09⁴⁴ was applied to find tandem repeats present in the scaffolds, and other repeats were also identified using homology-based approaches with Rebase version 21.06⁷⁰ which is a repository for known repeats. RepeatMasker version 4.0.6 (<http://www.repeatmasker.org>) was applied for DNA-level identification in combination with Rebase (**Supplementary Figures L4 A and B, L6 A and B**).

Gene prediction:

The prediction of Asiatic lion genes is described in four steps:

- a) **De novo prediction:** *de novo* prediction was performed on the repeat masked genome of Asiatic lion using AUGUSTUS version 3.1.0⁷¹ which is based on Hidden Markov Model (HMM). To perform the gene prediction, the training set of lion was created by using cat proteome and the lion assembly due to the unavailability of a trained data set. Scipio constructed gene structure from assembly of lion and proteins of *Felis catus* from UniProt⁷².

Those gene structures were used for the creation of training set. A hint file was also generated from the transcriptome data of lion and incorporated in to the gene prediction.

- b) **Homology-based prediction:** Homologous proteins of other species (cat, tiger, dog, human and mouse from NCBI) were mapped to the lion scaffolds using tBlastn (Blast 2.2.31) with an e-value cut off equal to 1E-3. The aligned scaffolds as well as its query proteins were then filtered and passed to Exonerate version 2.2.0⁷³ to search for accurate spliced alignments.
- c) **EST-based prediction:** ESTs of *Felis catus* were taken and aligned to the genome of Asiatic lion using BLAT-35x1 (identity \geq 0.90, coverage \geq 0.90)⁷⁴ to generate spliced alignments.
- d) **Integration evidence:** Source evidences generated from the three approaches mentioned above was integrated using EvidenceModeler-1.1.1⁷⁵ to produce consensus gene set. The summary evidences were created by performing blastn with all the predicted sequences from *ab initio*, homology and ESTs against EVM predicted CDS (**Supplementary Tables S18 and 19**).

Gene function annotation:

Gene functions were assigned according to the best alignment match with SwissProt and TrEMBL Mammalian database from UniProt⁷⁶. The annotation has been carried out using an in-house pipeline (CANoPI – Contig Annotator Pipeline). Briefly, we performed the following steps for annotation (**Supplementary Table 20**).

- Comparison with UniProt Mammalian database using BLASTp program
- Organism annotation
- Gene and protein annotation to the matched CDSs
- Gene ontology annotation
- Pathway annotation

Non-coding RNAs:

We detected four types of non-coding RNA in the Lion genome by searching databases using the whole genome sequence, tRNAscan-SE version 2.0⁴⁶ was performed on a SINE pre-masked genome to search for reliable tRNA positions. snRNA and miRNA were sought through a two-step method. After alignment with BLAST⁷⁷, INFERNAL 1.1.1 was used to search for putative sequences in the Rfam database release 12.1⁷⁸ (**Supplementary Table S17**).

GENE EVOLUTION:

Phylogenomic analysis:

a) Asiatic lion and genomes of other published mammalian genomes

A species tree depicting the inter-relationship of Asiatic lion and other mammalian genomes human (GRCh38.p9), mouse (GRCm38.p5), cheetah (aciJub1.0), tiger (PanTig 1.0), cat (Felis_catus-8.0), panda (aiMe1.0), dog (CanFam3.1) and opossum (monDom5.0) was reconstructed using concatenated amino acid sequences in the PHYLIP format using PhyML which applies BioNJ⁴⁹ which is a fast distance based method and helps in computing the initial tree quickly. Following that, fastNNI (nearest neighbor interchange) operations are applied for

tree optimization. The alternative NNI trees are evaluated and arranged according to their ML value. NNIs with maximum ML values are applied to the current tree and fastNNI continued till there is no scope for ML improvement. The LG substitution model was used. The resulting phylogenetic tree was visualized using TREEVIEW⁷⁹.

b) Asiatic lion with published genomes of other felids

The phylogenetic relationships of felids were reconstructed using four published genomes of felids (Asiatic lion, tiger, cat and cheetah) and panda and dog as outgroups. The tree was again reconstructed using the LG model of substitution in PhyML.

c) Divergence time estimation and rates of substitution

The divergence time for the eight mammalian genomes including Asiatic lion was estimated using amino acid sequences in the MCMCTREE implemented in the PAML package v 4.0⁵². The calibration time used for the human-mouse divergence (90Mya) was taken from the TimeTree database (<http://www.timetree.org>). The time of divergence between the Asiatic lion and Amur tiger was determined to be 10.5Mya (95% CI: 10.18-10.92Mya).

Orthologous protein families

The proteomes of eight mammals (Asiatic lion, Amur tiger, cat, dog, mouse, panda, opossum and human) were compared to examine the rate of protein evolution and conservation among protein orthologs using different databases meant for orthology prediction, OrthoMCL⁵⁴, InterPro⁸⁰ and Pfam⁵⁵.

OrthoMCL:

Among several databases and orthology prediction tools, the most sensitive and accurate is the OrthoMCL (utilizing the Markov Cluster Algorithm) program. OrthoMCL software is used to cluster proteins based on sequence similarity, using an all-against-all BLAST search of each species proteome, followed by normalization of inter-species differences, and Markov clustering. Based on sequence similarity and graph algorithm, OrthoMCL helps in distinguishing species proteomes into three separate categories; Orthologs, Co-orthologs and Paralogs, where the first two are speciation events (orthologs and co-orthologs) and the last one is the result of a duplication event (paralogs) within a genome (**Supplementary Table S21**).

The four felid proteomes of Asiatic lion, Amur tiger, domestic cat and Cheetah were compared to find Felidae-specific protein families. On combination of the four proteomes and removal of duplicates, a total of 126 protein families were found to be common to felids. Gene Ontology annotation could be done for 34 protein families (**Supplementary Table S24**). The PFam ID, InterPro ID and Ensembl ID were also determined for them. Comparisons with other genomes 17 genes found to be exclusive for lion and out of these 14 genes are annotated. The GO annotation, PFam ID, InterPro ID and Ensembl ID could be found for 9 protein families (**Supplementary Table S27**).

Expansion-contraction of protein families and gain-loss of proteins:

One method of assessing the phylogenetic relationships and deciphering the underlying evolutionary basis of certain significant metabolic phenotypes characterizing them is by evaluating the expansion-contraction of protein families or gain-loss of proteins among taxa.

Both these processes were examined by using the Wagner Parsimony algorithm of COUNT software computed by the AsymmetricWagner application⁶² in this study.

COUNT is a software package for the analysis of numerical profiles on a phylogeny. COUNT performs ancestral reconstruction, and infers family- and lineage-specific characteristics along the evolutionary tree. It implements popular methods employed in gene content analysis such as Dollo⁸¹ and Wagner parsimony⁸², Propensity for Gene Loss⁸³, as well as probabilistic methods involving a phylogenetic birth-and-death model⁸⁴. The asymmetric version of Wagner parsimony does differential penalization of gains and losses. Two input files were provided; the tree file (.txt format) and the Multiple Sequence Alignment (.tsv format) file. The output files were obtained for Dollo-Parsimony (DP), Wagner-Parsimony (WP) and PGL methods.

COUNT is for evolutionary analyses of phylogenetic profiles and other numerical characters and has been utilized here for estimating protein loss/gain and protein families' expansion/contraction. For ancestral reconstruction, asymmetric Wagner parsimony was applied and the gain penalty stringently set to 2.5. The lion protein families were further assigned to their InterPro IDs, GO and KEGG descriptions (**Supplementary Table S26**). PANTHER pathway analysis was also conducted (**Supplementary Table S27**).

Positively selected genes:

Detection of the strength and mode of natural selection on protein-coding genes is done by computing the ω statistic which is the ratio of rate of non-synonymous (dN) to synonymous (dS) substitutions⁸⁵. $\omega > 1$ when positive selection occurs along multiple codons of a gene and

throughout the entire phylogeny. But in most cases, positive selection is observed within a limited region of the protein and/or in some but not all the species in a phylogeny. Specific models have been developed to detect specific scenarios.

The statistic ω was estimated using the CODEML program of the Phylogenetic Analysis by Maximum Likelihood (PAML) package⁵⁶. dN and dS were calculated from orthologous protein-coding gene sequences obtained by alignment of annotated genomes of Asiatic lion, Amur tiger, cat, dog, mouse, panda, opossum and human. The orthologous gene sequences acted as input files for CODEML. The ortholog groups having more than one ortholog per species were removed retaining only 1:1 orthologs to avoid presence of paralogs that are a result of gene duplications. PRANK was used for multiple alignment of gene orthologs^{57,58}. The branch-site model of PAML was used in which ω can vary at specific sites of a gene sequence and branches of the phylogeny⁵⁹. A total of 240 lion protein-coding genes were determined as positively selected. The over-representation test was performed using the PANTHER database to identify gene families which are statistically over or under-represented in the list of positively selected genes but none of the protein families were found to be significantly over-represented. A pathway analysis was also carried out using PANTHER for this list of genes (**Supplementary Table S28**).

Genomic diversity:

The number of heterozygous SNVs per base pair of a genome has been used to comment on its genomic diversity in earlier papers on whole genomes^{13,14}. The lion raw reads were mapped to the cat genome to determine the number of heterozygous SNVs in the lion genome

(Supplementary Table S14). The rate of heterozygous SNVs for 9 felid genomes was compiled from Kim *et al.*¹³ and for lynx (Eurasian and Iberian) from Abascal *et al.*¹⁹ to visualize the position of Asiatic lion among felid genomic diversity **(Supplementary Table S29)**.

Male-specific genes in lion genome:

RNA-Seq data of five Asiatic lion individuals (3 males, 2 females) were mapped to the lion genome using STAR aligner⁸⁶. RNA-Seq by Expectation Maximization (RSEM) was used to quantify and perform differential analysis with $FDR < 0.5$ ⁸⁷. Gene expression values (expected counts) were calculated using RSEM, where expected counts are the number of paired reads for each gene. Genes with two-fold change were considered as differentially enriched in male samples. Genes that are one-fold more enriched in males were listed as male-specific genes **(Supplementary Table S30)**. Male-specific genes were selected based on a min value of 2 in PostFC which is the posterior fold change (condition 1 over condition 2) for a gene/transcript. It is defined as the ratio between posterior mean expression estimates of the gene/transcript for each condition.

GENOME EVOLUTION:

Segmental duplications:

The Whole Genome Assembly Comparison (WGAC) method was done to determine segmental duplications in the lion genome which is a BLAST based approach that performs an all-by-all comparison of assembled genomic sequence⁵⁹. The method, however, makes two assumptions: (1) the genome has been accurately assembled and (2) no under- or over-representation within

duplicated regions has occurred. A total of 6,174 recent segmental duplicated content of greater than 1kb length and more than 90% identity were found which accounts for 9.95Mb (0.4%) of the Asiatic lion genome. The low segmental duplication rate has been pointed out as an error in next generation sequencing assemblies by Cho *et al*¹⁴ in the Amur tiger assembly.

Chromosomal rearrangement:

- The domestic cat (*Felis_catus* 8.0) and lion chromosomes (this study) were assembled using the program *nucmer* available from MUMmer⁶⁴ which is a package used for rapid alignment of entire genomes.
- When one region/block of a chromosome happened to be mapped to distant cat chromosomal locations, they were noted as inter- or intra-chromosomal rearrangement events of the lion genome relative to the cat genome.
- First, similar regions in intra-chromosomal comparisons were identified (Supplementary Figure L11). Here regions from same locations were discarded first. Then smaller regions which are less than 5Kb were discarded. Finally, the regions those are mapped to distant locations of the cat genome were listed.
- For inter-chromosomal comparisons, we considered all regions which are greater than 5Kb in length for rearrangement. Chromosomal rearrangement summary is shown in Supplementary Table S 35.

Additional Supplementary References:

65. Grabherr, M.G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **29**,644 (2011).
66. Langmead, B. and Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357 (2012).
67. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
68. Li, H. and Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
69. Rice, P., Longden, I. and Bleasby, A. EMBOSS: the European molecular biology open software suite. *Trends in Genetics* **16**, 276-277 (2000).
70. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, **27**, 573-580 (1999).
71. Bao, W., Kojima, K.K. and Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, **6**, 1(2015).
72. Stanke, M., Diekhans, M., Baertsch, R. and Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637-644 (2008).
73. Keller, O., Odrionitz, F., Stanke, M., Kollmar, M. and Waack, S. Scipio: using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. *BMC Bioinformatics* **9**, 278 (2008).
74. Slater, G.S.C. and Birney, E., 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
75. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Research*, **12**, 656-664 (2002).
76. Haas, B.J.*et al.* Automated eukaryotic gene structure annotation using EvidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology*, **9**, 1 (2008).
77. Bairoch, A. and Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, **28**, 45-48 (2000).
78. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. Basic local alignment search tool. *Journal Molecular Biology*, **215**, 403-410 (1990).

79. Nawrocki, E.P. and Eddy, S.R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933-2935 (2013).
80. Page R. D. M., "Tree View: An application to display phylogenetic trees on personal computers," *Bioinformatics* **12**, 357–358 (1996).
81. Hunter, S. *et al.* InterPro: the integrative protein signature database. *Nucleic Acids Research*, **37**(suppl_1), D211-D215 (2008).
82. Farris, J.S. Phylogenetic analysis under Dollo's Law. *Systematic Biology*, **26**, 77-88 (1977).
83. Farris, J.S. Methods for computing Wagner trees. *Systematic Biology*, **19**, 83-92 (1970).
84. Krylov, D.M., Wolf, Y.I., Rogozin, I.B. and Koonin, E.V. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Research*, **13**, 2229-2235 (2003).
85. Kendall, D.G. Stochastic processes and population growth. *Journal of the Royal Statistical Society. Series B (Methodological)*, **11**, 230-282 (1949).
86. Kryazhimskiy, S. and Plotkin, J.B. The population genetics of dN/dS. *PLoS Genetics*, **4**, p.e1000304 (2008).
87. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
88. Li, B. and Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).