Comparative sequence analyses of genome and transcriptome reveal novel transcripts and variants in the Asian elephant *Elephas maximus*

Puli Chandramouli Reddy¹,†, Ishani Sinha²,†, Ashwin Kelkar¹,†, Farhat Habib¹,†, Saurabh J. Pradhan¹, Raman Sukumar²,† and Sanjeev Galande¹,*,‡

¹Centre of Excellence in Epigenetics, Indian Institute of Science Education and Research, Pashan, Pune 411 008, India
²Centre for Ecological Sciences, Indian Institute of Science, Bangalore 560 012, India

*Corresponding author (Fax, 91-20-25865086; Email, sanjeev@iiserpune.ac.in)
†These authors contributed equally to the work.
‡Shared senior authorship.

The Asian elephant *Elephas maximus* and the African elephant *Loxodonta africana* that diverged 5–7 million years ago exhibit differences in their physiology, behaviour and morphology. A comparative genomics approach would be useful and necessary for evolutionary and functional genetic studies of elephants. We performed sequencing of *E. maximus* and map to *L. africana* at ~15X coverage. Through comparative sequence analyses, we have identified Asian elephant specific homozygous, non-synonymous single nucleotide variants (SNVs) that map to 1514 protein coding genes, many of which are involved in olfaction. We also present the first report of a high-coverage transcriptome sequence in *E. maximus* from peripheral blood lymphocytes. We have identified 103 novel protein coding transcripts and 66-long non-coding (lnc)RNAs. We also report the presence of 181 protein domains unique to elephants when compared to other Afrotheria species. Each of these findings can be further investigated to gain a better understanding of functional differences unique to elephant species, as well as those unique to elephantids in comparison with other mammals. This work therefore provides a valuable resource to explore the immense research potential of comparative analyses of transcriptome and genome sequences in the Asian elephant.


1. Introduction

Elephants are the largest living land mammals in the world, belonging to the family Elephantidae and order Proboscidea, which evolved 60 million years ago (Shoshani and Tassy 1996; Sukumar 2003). Two elephant species that are well known are the Asian elephant, *Elephas maximus*, and the African elephant, *Loxodonta africana*. Morphological as well as molecular evidence suggests that the African elephant comprises two distinct species, the savannah elephant *L. africana* and the forest elephant *L. cyclotis* (Roca et al. 2001; Rohland et al. 2007, 2010). Together, they comprise the only living representatives of the order Proboscidea, which includes 175 species (Shoshani and Tassy 2005). Proboscideans in general and members of the family Elephantidae in particular, including the recently extinct mammoths (*Mammuthus* spp.), shared several unique

Keywords. Asian elephant; comparative genomics; gene prediction; transcriptome

Supplementary materials pertaining to this article are available on the *Journal of Biosciences* Website at http://www.ias.ac.in/jbiosci/dec2015/supp/Reddy.pdf

http://www.ias.ac.in/jbiosci J. Biosci. 40(5), December 2015, 891–907, © Indian Academy of Sciences 891

Published online: 4 December 2015
features common to the extant elephants, such as large body size, proboscis that formed long muscular trunks and unique dental characteristics (Osborn and Percy 1936). One of the distinctive features of present-day elephants, the tusks, were also present during the evolution of early Proboscidea (Osborn and Percy 1936). Elephants are also known for their exceptional olfactory sense and chemical communication (Rasmussen and Krishnamurthy 2000; Shoshani et al. 2006), memory and intelligence (Bates et al. 2008; Byrne et al. 2009; Foerder et al. 2011), unique vocal communication (Payne 2003; Nair et al. 2009; Stoeger and de Silva 2014) and social organization (Moss 1988; de Silva and Wittemyer 2012).

The closest living relatives of elephants are the hyraxes (order Hyracoidea) and sea cows (Sirenia), followed by aardvarks (Tubulidentata), elephant shrews (Macroscelidea), tenrecs and golden moles (Afrosoricida), which together comprise the superorder Afrotheria (Springer et al. 1997). Despite the radical differences in body size and visible morphological features, members of this group share several developmental characteristics such as a primitive placenta (Carter et al. 2006), increased number of thoracolumbar vertebrae (Sánchez-Villagra et al. 2007), late eruption of permanent teeth (Asher and Lehmann 2008), non-descended male gonads (Werdelin and Nilsson 1999), etc. Molecular and phylogenetic analyses have conclusively proved the monophyletic origin of this group (reviewed in Tabuce et al. 2008; Swartman and Stanyon 2012).

A number of investigations on elephants thus far have been focused on their ecology, behaviour, phylogeny, population genetics and conservation. However, since the early 1990s, molecular analyses for resolving phylogenies and population genetics in elephants have been carried out by sequencing a few mitochondrial regions or nuclear genes and microsatellite loci. With the advent of high-throughput sequencing technologies, sequence data from hundreds of loci were used to resolve fine-scale phylogenetic relationships between the African elephants, the Asian elephant and the woolly mammoth (Rohland et al. 2010). This study indicated that the Asian elephant was the closest relative of the woolly mammoth, and that the forest and savannah African elephants were at least as diverged from each other as the Asian elephant from the woolly mammoth. Another study used shotgun and RAD sequencing to identify de novo markers for a genetically isolated population of elephants in Borneo with extremely low levels of genetic diversity, which could be used in population genetic studies (Sharma et al. 2012).

The first version of the genome assembly was made available in May 2005, and the third version used for all analysis was released in June 2009 (https://www.broadinstitute.org/scientific-community/science/projects/mammals-models/elephant/elephant-genome-project). Subsequently, the whole genome of the Asian elephant was sequenced at low coverage (Dastjerdi et al. 2014). Recently, high coverage genome sequences of three Asian elephants along with two mammoth genomes have been published (Lynch et al. 2015). High-coverage whole genome sequences are also available for several other Afrotherians, but remain poorly analysed.

A few studies on comparative genomics have also been carried out in elephants. One of the first comparative genomic studies compared the placental transcriptome of the African elephant with that of other eutherian mammals (Hou et al. 2012). This study identified thousands of commonly expressed genes, as well as lineage-specific genes that tracked the evolution of placental characteristics across the mammalian phylogenetic tree. The fibroblast transcriptome of African elephants was also sequenced and used in a study to trace the evolution of the Y-chromosome in mammals (Cortez et al. 2014). A phylogeny-based comparison of olfactory genes across 13 mammals revealed an unusually large repertoire of functional olfactory receptor genes among the African elephants when compared to other mammals (Niimura et al. 2014) and substantiated the well-documented observations that elephants have an exceptional sense of smell. In another study, the molecular bases of cold adaptation in mammoths was carried out by comparative sequence analyses between whole genome sequences of Asian elephants and mammoths against the African elephant reference genome and identification and analysis of mammoth specific SNVs that fall within protein coding regions (Lynch et al. 2015). Two recent studies have elucidated a possible mechanism for the reduced occurrence of cancer in Asian and African elephants (Abegglen et al. 2015; Sulak et al. 2015) compared to other mammals including humans. Comparative genome and phylogenetic analyses showed that the elephantid lineage had evolved multiple copies of TP53, a tumour suppressor gene that is present as a single copy in other mammals. Some of the extra copies of TP53 were shown to be expressed, and elephant cells showed a higher sensitivity to DNA-damaging radiation, and an increased induction of apoptosis when compared to human cells.

In this study, we have carried out whole genome and transcriptome next generation sequencing and performed a comparative genomics analysis with other elephantid sequences available from public databases. We were able to uncover several interesting leads that can be investigated further and will significantly contribute to building genomic resources for functional genetic research in elephants in the future. We identified Asian elephant specific homozygous single nucleotide variants (SNVs), many of which mapped within protein coding regions and resulted in nonsynonymous changes. We used our transcriptome data to improve the functional annotation of predicted genes in the L. africana reference genome. We identified potentially novel transcripts from expressed transcript fragments in our dataset that did not map to any known or predicted genes in
the *L. africana* genome. We also analysed the unmapped transcript fragments to identify lncRNAs which might be involved in tissue and cell-type-specific functions. In addition, we compared the domain architectures of elephant proteome sequences with those of three other Afrotheria species and identified domain architectures that are overrepresented in elephants.

2. Methods

2.1 Sample collection, extraction of nucleic acids and next-generation sequencing

Peripheral blood was collected from the ear vein of a captive male Asian elephant Jayaprakash from the Karnataka Forest Department’s elephant camp at Bandipur National Park, Karnataka, India (permit details provided in Acknowledgments). Venous blood was collected in PAXgene blood RNA Tubes (Cat no. 762165 from PreAnalytiX) and stored at −20°C.

White blood cells (WBCs) were extracted from the blood by lysing the red blood cells (RBCs) with 0.7 M KCl. Genomic DNA from total lymphocytes was isolated from peripheral blood (WBCs) using PureLink genomic DNA isolation kit. DNA integrity was assessed by electrophoresis on 0.8% agarose gel. Total RNA from lymphocytes was isolated from peripheral blood (WBCs) using Trizol and the quality of RNA was analysed using Bioanalyzer 2100 RNA 6000 nano Chip (Agilent). Five μg of high quality genomic DNA and 1 μg RNA samples with RIN value of 7.2 were used for library preparation and sequencing.

For whole genome sequencing, a paired end library with 400–500 bp insert size and a mate-paired library with 2–3 kb insert size were prepared. The paired end (PE) library was prepared using Illumina Truseq library kit and 2×100 bp sequencing reads were obtained using Illumina HiSeq 2500, for approximately 15X coverage. The mate pair library was prepared using the Illumina Nextera Mate-Pair library kit and sequenced to obtain 2×250 bp reads for 5X coverage. Bioanalyzer plots were used at every step to assess fragmentation sizes and final library sizes. Prior to loading on the sequencer, the quantity of the library was estimated using picogreen assay (Thermo Fisher) and quantitative PCR (qPCR).

For the transcriptome sequencing, the RNA-seq library was prepared using TruSeq RNA Sample Prep Kits (Illumina). Briefly, poly-A containing mRNA molecules were enriched using poly-T oligo-attached magnetic beads. mRNA was fragmented to obtain desired size using divalent cations under elevated temperature. The cleaved RNA fragments were used to synthesize first strand cDNA using reverse transcriptase and random primers followed by second strand cDNA synthesis using DNA Polymerase I and RNase H. These cDNA fragments were end repaired, and adapters were ligated. The library was further enriched with PCR to create the final cDNA library and sequenced using 2×100 bp PE sequencing with Illumina HiSeq 2500. Bioanalyzer plots were used at every step to assess mRNA quality, enrichment success, fragmentation sizes, and final library sizes. Prior to loading on the sequencer, the quantity of the library was estimated using Qubit fluorimeter (Thermo Fisher) and qPCR.

2.2 Genome sequence mapping

Raw FASTQ reads were checked using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads had excellent overall quality for both the paired-end and mate pair runs. Read mapping was performed using the Burrows-Wheeler Aligner (BWA) (Li and Durbin 2009) (bwa mem algorithm was used) with default parameters for the paired-end reads and with -L 80,80 option for mate-pair reads to suppress clipped alignments using the *L. africana* (loxAfr3) genome sequence as reference. The loxAfr3 genome was obtained from the Broad Institute. Both the mate-pair and paired end runs achieved a greater than 98% alignment rate, though ~9% of the reads were clipped for paired end alignments and 82% of the reads were clipped for the mate-pair sequences. The high rate of clipping for mate-pair sequences indicates potential issues in downstream analysis; hence we focused on the paired end run for most of the further analysis. The raw genomic DNA sequence data files were deposited at NCBI SRA under BioProject ID PRJNA301482, BioSample ID SAMN04260428 and run ID SRR2912975.

2.3 Genome rearrangement analyses

To visualize genome rearrangements using Circos (Krzywinski et al. 2009) we used uniquely aligning paired end reads whose ends aligned onto different chromosomes (samtools view -F 1294 -q 1), we then binned the reads into 1 kb bins and only considered bins where more than 25 reads were aligning into bins in either end as reliable links to weed out chimeric reads or misalignments. This was performed using shell and awk scripts. To visualize the *E. maximus* genome, all scaffolds were laid end to end in order of decreasing length with no gaps, and links identified were connected via arcs.

2.4 Variant calling and identification of Asian elephant specific SNVs

We carried out variant calling of our Asian elephant genome sequence using African elephant genome as reference. The Samtools/vcftools variant calling pipeline was used with default parameters except for -D250 parameter (ignoring
regions with alignment depth greater than 250 as was followed by Lynch et al. (2015) and combining both the paired-end and mate-pair alignments.

We also followed a modified Galaxy pipeline (https://usegalaxy.org/r/woolly-mammoth) and a dataset of Asian elephant and mammoth variants, ‘mammoth SNPs’ from Lynch et al. (2015) to filter the subset of homozygous variants present in all the three Asian elephant sequences but not in the two mammoth sequences. We compared the homozygous variants from our data with the subset of Asian elephant specific homozygous variants obtained from the Lynch et al. (2015) dataset to determine the extent of overlap. We then joined the subset of Asian elephant specific SNVs with the ‘mammoth coding SNPs’ dataset (also from Lynch et al. 2015) to identify the Asian elephant specific variants that mapped to protein coding regions and resulted in non-synonymous mutations. We retrieved the FASTA protein sequences of all the genes in this dataset from the *L. africana* genome and carried out a BLAST search using default parameters to identify the top hits in other mammals. All the unique Uniprot IDs for the top mammalian homologs were analysed for associations with known gene ontology using DAVID (da Huang et al. 2009; Huang et al. 2009) and the BiNGO app in Cytoscape (Shannon et al. 2003; Maere et al. 2005). The list of biological process GO terms from the DAVID output were analysed for semantic similarity with REVIGO (Supek et al. 2011) and the overlap between the gene lists of significant GO terms was determined using the UpSetR package (Gehlenborg 2015) in RStudio (R CoreTeam 2014).

2.5 Transcriptome analysis

We performed reference based transcriptome assembly using the Tuxedo pipeline (Trapnell et al. 2012) using the *L. africana* genome sequence as reference with annotations from NCBI (http://www.ncbi.nlm.nih.gov/genome/224?genome_assembly_id=28577). We first executed TopHat with transcriptome reads and obtained alignments for all reads. We downloaded the transcriptome sequences from two male African elephant fibroblasts available in GEO dataset (GSM1227965 and GSM1278046) (Cortez et al. 2014) and performed similar analyses for comparison with our transcriptome. Count of reads aligned to mRNA feature were calculated using MultiCov option in BEDtools (Quinlan and Hall 2010) using *L. africana* GFF as reference. Next, the read count data of all the three transcriptomes were modified by adding 1 to all counts to include genes not expressed in one of the samples and to avoid artifacts in log transformation. The data were then log transformed and filtered to include data with a minimum of 5 counts per mRNA feature, and only genes with a minimum difference of 1 log transformed value, among the samples were selected. Subsequently, the genes were centered based on log values and hierarchical clustering was carried out based on calculated Pearson correlation and displayed based on centroid of the clusters using Cluster3 (de Hoon et al. 2004). A cluster of highly expressed genes from the WBC transcriptome of this study compared to the two fibroblast transcriptomes was further analysed for associated functions.

In addition to the cluster analysis, we performed comparative expression analysis using Cuffdiff (Trapnell et al. 2010) and visualized the data by drawing a volcano plot with CummeRbund in R programming language (Goff et al. 2012). The raw RNA sequence data files were deposited at NCBI SRA under BioProject ID- PRJNA301482, BioSample ID- SAMN04247652 and run ID SRR2911072.

Functional enrichment analysis of selected cluster of genes selected was carried out by BiNGO app in Cytoscape (Shannon et al. 2003; Maere et al. 2005). A hypergeometric test was performed to calculate the overrepresented GO terms with Bonferroni Family-Wise Error Rate (FWER) correction. Ontology terms belonging to Biological Process with less than 0.05 p-value were selected and mapped hierarchically based on parent and child relation.

2.6 Annotation of predicted protein models

Functional annotation of hybrid GFF (Gene Feature File) generated by combining reference from *L. africana* available at NCBI and novel transcripts identified from current study was used. Here, InterProScan-5 pipeline (Jones et al. 2014) was used according to the software manual against ProDom, ProSiteProfiles, PRINTS, Pfam and TIGRFAM databases, associated GO (gene ontology) terms were extracted and a custom GO Annotation File (GAF) was built and used for further analysis.

2.7 Identification of novel transcripts

We used the Tuxedo pipeline to identify novel transcripts (Trapnell et al. 2012) that were not annotated in L. africana reference genome. First, we used cufflinks to generate transcript fragments (transfags) and a gene transfer format (GTF) file of the same using TopHat aligned reads using *L. africana* reference GTF from NCBI. The output GTF file from cufflinks was used for comparing reference GTF and Cufflinks generated GTF using Cuffcompare. From the Cuffcompare output file, transfags aligned uniquely to regions marked as intergenic in the *L. africana* reference genome were considered as putatively novel transcripts. These were further filtered by Transdecoder for identifying potential novel transcripts coding for peptides. These coding
peptides were annotated using InterProscan-5 to determine the associated function.

Gene ontology analysis of novel genes identified in this study was done using REVIGO. Here, list of GO terms associated with novel transcripts were used with 0.9 allowed similarity using SimRel option for semantic calculation against UniProt database. Results of REVIGO are represented as treemap.

2.8 Identification of lncRNAs

We utilized the tools TransDecoder (Haas et al. 2013) and Coding potential calculator (CPC) (Kong et al. 2007) which classify RNAs based on presence or absence of protein coding domains and codon usage frequencies to the 3904 unannotated transcripts.

The transcripts that were predicted to be ncRNAs by both the tools were then further analysed using RFam database to identify any possible homology to known ncRNAs. The annotations were based on a Covariance model and/or Hidden Markov Model that denote a specific signature representing a specific lncRNA. These models are generated using all the known transcripts for each reported and annotated ncRNA. These predictive models were used to compare against our sequences of interest, for identification of any significant match with a known ncRNA.

2.9 Comparative domain composition analysis

A comparative analysis of protein domain architectures was performed by mapping domain HMM (hidden markov model) models of reference proteomes of three Afrotheria species: tenrec (Echinops telfairi, http://www.ncbi.nlm.nih.gov/genome/234), manatee (Trichechus manatus latirostris, http://www.ncbi.nlm.nih.gov/genome/24630), and aardvark (Orycteropus afer, http://www.ncbi.nlm.nih.gov/genome/11840) with the L. africana refseq proteome build from the current analysis and refseq available in NCBI database (http://www.ncbi.nlm.nih.gov/genome/224). Here, initially hmmscan available in HMMER3 suit (Johnson et al. 2010) was performed with default parameters with all the proteomes and proteins. Unique domain sets were extracted from each hmmscan outfile and all lists were compared using VENNY tool (Oliveros 2007). Based on this analysis, domains unique to the L. africana sequence were selected for further characterization. Different levels of domain architectures based on their numbers were calculated for understanding the group specific changes within the selected species.

Generic commands and scripts used in this study are given in supplementary file 2.

3. Results

3.1 Genome sequence analysis

A total of 472,003,946 reads were obtained from the paired end sequencing and in 95.63% of the reads, both reads in the pair were mapped, amounting to a mean coverage of 13.9X with a mean mapping quality of 28.89. A total of 88,920,076 reads were sequenced from the mate-pair library and 99.14% of the reads had both reads in the pair mapped, amounting to a mean coverage of 4.15X with a mean mapping quality of 25.45 (supplementary figure 1A and B). The BWA tool was able to find alignments for >99% of the mate-pair reads and >95% of the paired end reads, though multi-part alignments were obtained for many reads as is possible in the presence of structural variations, gene fusion or reference mis-assembly. This is expected as the reference genome is from a different genus.

3.2 Genome rearrangement analyses

Circos plot of paired end reads from E. maximus that aligned on different scaffolds of the L. africana genome indicate putative genomic rearrangements or scaffold interconnections in the Asian elephant as compared to the African elephant genome (figure 1). To simplify the figure and show only reliable rearrangement sites, we isolated uniquely aligning reads that aligned on different scaffolds. The reads were then binned in 1Kb bins, and only regions where more than 25 reads aligned within the same bin on both ends were connected with an arc.

3.3 Variant analysis

We identified a total of 28.29 million SNVs in our genome sequence when compared to the L. africana genome, which included indels, heterozygous as well as homozygous variants. Of these, about 21.7 million SNVs were homozygous. The size of the African elephant genome is 3.1 Gb (http://www.ncbi.nlm.nih.gov/genome/224?genome_assembly_id=28577) and the Asian elephant genome is assumed to be of similar size (Dastjerdi et al. 2014). The average SNV density was 8.85 SNVs per kb, with most of the regions containing about 20 SNVs per kb (supplementary figure 2). In parallel, we identified ~1.9 million homozygous Asian elephant specific SNVs defined as variants (compared to the African elephant reference genome) that were homozygous across all the three Asian elephant sequences in the Lynch et al. (2015) dataset but absent from both the mammoth sequences and found a 95% overlap between our dataset and the Lynch et al. Asian elephant specific homozygous SNVs. Of the Asian elephant specific SNVs 7155 SNVs mapped to protein coding regions of the L. africana

J. Biosci. 40(5), December 2015
genome, and 1859 of them resulted in non-synonymous changes in the coding sequences. The list of all Asian elephant specific non-synonymous mutations is provided in supplementary table 1. The 1859 non-synonymous variants mapped to 1514 unique protein coding genes, of which 1506 were mapped to genomes of mammals by BLAST search, while 8 returned no significant hits. FASTA sequences of the 1506 genes were used to retrieve 1367 unique Uniprot IDs in higher mammals 1259 of which were mapped to GO terms in the DAVID annotation tool. We focused on 81 biological

Figure 1. Circos plot depicting paired end reads from *E. maximus* that aligned on different scaffolds of the *L. africana* genome. The *L. africana* genome (loxAfr3) is laid out clockwise starting from the 12 o’clock position with scaffolds in order of decreasing length. As there are over 2300 scaffolds in loxAfr3, gaps between the scaffolds are not shown. The numbers on the circumference refer to Gigabases with the elephants genome being ~3.1Gb. This figure shows clusters of paired end reads where the two ends of the pairs align on different scaffolds. A large number of these happen in the smaller scaffolds which could potentially be connected into larger scaffolds providing a more complete genome for *E. maximus* using the *L. africana* genome as a reference. The links have been coloured by scaffold position to make differentiating them easier.
process GO terms (supplementary table 2) and semantic analyses with REVIGO indicated three terms with highly significant \( \log_{10} p \)-values: sensory perception of smell, cell surface receptor protein signalling pathway and G-protein coupled receptor protein signalling pathway (figure 2A). Applying a stringent Bonferroni corrected cutoff of 0.05 yielded 7 significant biological processes: GO:0007608 (sensory perception of smell), GO:0007606 (sensory perception of chemical stimulus), GO:0007600 (sensory perception), GO:0050890 (cognition), GO:0050877 (neurological system process), GO:0007186 (G-protein coupled receptor protein signalling pathway), GO:0007166 (cell surface receptor linked signal transduction) (table 1). However, there was a significant overlap between these categories. The lists of genes categorized in the GO terms sensory perception of smell (GO:0007608) and sensory perception of chemical stimulus (GO:0007606) almost completely overlapped with each other, and were subsets of the GO term sensory perception GO:0007600 which in turn was found to overlap significantly with cognition (GO:0050890). Moreover, sensory perception and cognition were found to be subsets of neurological system process (GO:0050877) which overlapped significantly with the two remaining GO terms G-protein coupled receptor protein signalling pathway (GO:0007186) and cell surface receptor linked signal transduction (GO:0007166) (figure 2B). Significantly, a set of 113 olfactory receptor genes (supplementary table 3) was common to all the seven GO categories, while 38 genes were unique to cell surface receptor linked signal transduction and 11 were unique to neurological system process. The GO categories under cellular processes and molecular function indicated similar trends with enrichment of functions related to olfactory receptor activity (GO:0004984), plasma membrane (GO:0005886) and integral to membrane (GO:0016021). An independent analysis of functional enrichment using BinGO in Cytoscape also yielded similar results with an enrichment of olfactory receptor activity (supplementary figure 3).

### Table 1. List of the top ten biological process GO terms

<table>
<thead>
<tr>
<th>GO ID</th>
<th>Term</th>
<th>Count</th>
<th>%</th>
<th>( p )-Value</th>
<th>Fold Enrichment</th>
<th>Bonferroni</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0007608</td>
<td>sensory perception of smell</td>
<td>113</td>
<td>8.98</td>
<td>4.21784E−48</td>
<td>4.95</td>
<td>9.36782E−45</td>
</tr>
<tr>
<td>GO:0007606</td>
<td>sensory perception of chemical stimulus</td>
<td>114</td>
<td>9.05</td>
<td>4.28044E−44</td>
<td>4.51</td>
<td>9.50687E−41</td>
</tr>
<tr>
<td>GO:0007600</td>
<td>sensory perception</td>
<td>136</td>
<td>10.80</td>
<td>6.0228E−35</td>
<td>3.17</td>
<td>1.33766E−31</td>
</tr>
<tr>
<td>GO:0050890</td>
<td>cognition</td>
<td>139</td>
<td>11.04</td>
<td>2.61192E−31</td>
<td>2.89</td>
<td>5.80108E−28</td>
</tr>
<tr>
<td>GO:0050877</td>
<td>neurological system process</td>
<td>153</td>
<td>12.15</td>
<td>2.06571E−25</td>
<td>2.39</td>
<td>4.58795E−22</td>
</tr>
<tr>
<td>GO:0007186</td>
<td>G-protein coupled receptor protein signalling pathway</td>
<td>141</td>
<td>11.20</td>
<td>5.25912E−23</td>
<td>2.37</td>
<td>1.16805E−19</td>
</tr>
<tr>
<td>GO:0007166</td>
<td>cell surface receptor linked signal transduction</td>
<td>180</td>
<td>14.30</td>
<td>3.77473E−17</td>
<td>1.83</td>
<td>8.38367E−14</td>
</tr>
<tr>
<td>GO:0000279</td>
<td>M phase</td>
<td>38</td>
<td>3.02</td>
<td>1.16E−05</td>
<td>2.18</td>
<td>0.025339209</td>
</tr>
<tr>
<td>GO:0007018</td>
<td>Microtubule-based movement</td>
<td>19</td>
<td>1.51</td>
<td>2.49E−05</td>
<td>3.18</td>
<td>0.053890643</td>
</tr>
<tr>
<td>GO:0022403</td>
<td>cell cycle phase</td>
<td>42</td>
<td>3.34</td>
<td>7.92E−05</td>
<td>1.92</td>
<td>0.161278645</td>
</tr>
</tbody>
</table>

Only the top 7 GO terms marked in green were used for further analyses.
Figure 2. Gene ontology terms (biological processes) enriched in the Asian elephant specific non-synonymous SNVs. (A) Semantic analyses of all biological processes GO terms identified in our dataset in REVIGO. Colour scale is the log 10 of p-value for each GO term. Three terms: sensory perception of smell, G-protein-coupled receptor signalling pathway and cell surface receptor signalling pathway are significantly enriched. (B) Analysis of overlap between the 7 GO term gene sets with significant Bonferroni corrected p values using UpSetR. Sensory perception of smell (GO:0007608), Sensory perception of chemical stimulus (GO:0007606), Cognition (GO:0050890), G-protein coupled receptor protein signalling pathway (GO:0007186), Neurological system process (GO:0050877) and Cell surface receptor linked signal transduction (GO:0007166).
3.4 Transcriptome analysis

Out of 35,052,348 reads from RNA sequencing data, 74.1% reads were aligned to L. africana reference genome suggesting a good coverage of the total transcriptome. Next, to evaluate the correlation in expression levels among different samples and to identify the genes unique to WBCs, a cluster analysis with the two transcriptome sequences of African elephant male fibroblasts was performed. We focused on genes that were different in their log transformed expression values by a minimum of 1 between the two fibroblast datasets and our WBC dataset. As expected, the two fibroblast transcriptomes clustered together, while our WBC transcriptome occupied a separate node. We could identify 2 distinct clusters of genes, one that was up-regulated and another that was down-regulated in the WBCs when compared to the two fibroblast transcriptomes (figure 3A). Upon annotation using NCSeq for L. africana containing 29784 proteins, we could annotate 28638 protein coding genes using InterProScan. We obtained 3904 novel transcript fragments (supplementary file 1) that did not map to the L. africana reference genome annotation file in our transcriptome analysis. A custom annotation file was created using our annotations for functional enrichment analysis. A more detailed analysis showed a sub-cluster of 458 genes (Supplementary Table 4) that were highly expressed only in our WBC transcriptome, but not in the fibroblasts. Functional enrichment using our custom annotation files indicated functions associated with the immune system (figure 3B). This suggests good representation of the transcriptome specific to WBCs in our data. An additional differential expression analysis was performed using Cuffdiff tool (part of Tuxedo suite). In this analysis 9019 genes showed significant differential expression in comparison with the fibroblast transcriptome (figure 3C). In addition, 321 of the 458 genes selected as sub-cluster unique to WBC from cluster analysis were also found significantly upregulated in WBC by Cuffdiff analysis.

We identified 3,904 transcripts aligning to the intergenic regions that do not have any gene annotation associated with it. Out of these 1,458 could potentially code for peptides. We were able to annotate 103 of these 1,458 predicted peptide models using InterProScan. Most of these 103 novel proteins (supplementary table 5) were found to be associated with immune-related functions (figure 4).

3.5 Identification of IncRNAs

Since we used gene prediction models based on alternative splicing patterns to align the transcripts in our dataset to the L. africana assembly, and a majority of the IncRNAs have a maximum of 2 exons and do not undergo alternative splicing, it was possible that at least some of the 3,904 transcripts that were un-annotated could be IncRNAs which could not be mapped to predicted gene models. Thus an independent analysis of these 3,904 transcripts was performed to explore the possibility of them being IncRNAs. Since IncRNAs have unique sequence features that separate them from all the protein coding genes, any combination of the following 3 parameters could be used to identify potential IncRNAs (Dinger et al. 2008): (1) Nucleotide frequencies of protein coding regions are dictated by non random codon usage, (2) Protein coding regions typically contain known protein coding domains, thus sequence analysis against databases like Pfam can be used to identify coding regions, (3) Protein coding regions tend to have sequence similarities with protein sequence databases as conservation of amino acid sequence exerts a much higher selection constraint on protein coding regions as opposed to non-coding genes.

TransDecoder predicted 1,503 transcripts as potential ncRNAs while Coding Potential Calculator (CPC) predicted 2904. Of these, 66 ncRNAs were identified based on homology to known IncRNAs that are expressed in other mammals with an e-value 0.01 (supplementary table 6). Our dataset includes some extremely well studied IncRNAs (summarized in table 2) such as PVT1_3, DLEU2_1, ZNRD-AS1_1 that were identified with a high significance. PVT1_3 and DLEU2_1 are oncogenes and span translocation breakpoints that are related to Burkitt’s lymphoma and acute B cell lymphocytic leukemia respectively while ZNRD-AS1_1 is expressed from the MHC I locus specifically. Our dataset also contains well-characterized IncRNAs MALAT1, NEAT1 that are universally expressed and highly conserved in mammals with respect to expression pattern and function. The highly significant matches to these IncRNAs yield credibility to our dataset as being a good representation of the transcribed IncRNAs in E. maximus. In addition to these ‘benchmark’ IncRNAs, we also detected other transcripts that are reported to be involved in diverse functions, ranging from circadian rhythms to cellular apoptosis (Table II).

3.6 Domain architecture

We compared the domain architectures of the proteomes of L. africana with three other species of Afrotheria, the tenrec, manatee and aardvark. The total number of predicted proteins in these species ranges from ~23,000 to ~29,000 of which 5,936 proteins shared similar domain architectures. We observed a significant gain in the number of proteins in most of the architectures in the African elephant. Specifically, proteins with 11 to 20 domains have doubled in number in the African elephant in comparison to tenrecs (E. telfairi) (figure 5A.) (supplementary table 7). Moreover, 181 domains were found to be unique to the elephant. 

J. Biosci. 40(5), December 2015
Figure 3. Gene expression analysis of the Asian elephant WBC transcriptome. (A) Hierarchical clustering of genes based on their expression from two African elephant fibroblast transcriptomes and the Asian elephant WBC transcriptome from the present study. Clustering was performed by Cluster3 after log transforming the counts of aligned reads to transcripts. Scale bar represents all values normalized to the maximum read count and red coloured transcripts are upregulated, green coloured transcripts are down regulated whereas black colour transcripts show no change across the samples. (B) Functional enrichment analysis of selected cluster of genes upregulated in the WBC transcriptome. Enrichment analysis of the highlighted cluster of transcripts was carried out by BiNGO tool in Cytoscape using custom GO Annotation File (GAF) made from InteProScan-5 outfile. Scale bar shows the p-value score. Size of the circles is proportional to number of genes with the corresponding GO term. (C) Volcano plot displaying differentially expressed genes between African elephant fibroblast transcriptome (q1) and Asian elephant WBC transcriptome (q2). Differential expression analysis was carried out by cuffdiff (part of the Tuxedo suite) and results were plotted using cummerRbund package using R programming language. Each dot represents a gene and red coloured dots denote significantly differentially regulated genes.
(figure 5B) and only 33 of these domains had corresponding related GO terms (supplementary table 8).

4. Discussion

Functional genetic studies among ‘non-model’ or wild animals, such as elephants, is often severely restricted because of limitations such as access to samples, problems with establishing proper controls and treatments, biological variability, inability to carry out genetic manipulations, to name a few. Many such ‘non-model’ organisms may hold the key to answering pertinent questions with respect to the evolution of specific adaptive traits. While it is not possible to create laboratory models that would cover the entire spectrum of biological diversity, the recent availability of whole genome and transcriptome sequences of several organisms have allowed the use of comparative analyses to provide valuable insights into the evolution of phylogenetic groups as well as the evolutionary history of specific traits (Mikkelsen et al. 2007; Ellegren 2008; Warren et al. 2008; Lindblad-Toh et al. 2011; Parker et al. 2013; Keane et al. 2015). In particular, comparative transcriptomics not only helps to identify novel transcripts and novel gene expression patterns, but also aids in the discovery and characterization of non-coding RNAs that are widely expressed and perform many important functions (Trapnell et al. 2010; Xie et al. 2014). Of particular interest are the long non-coding (lnc)RNAs that are longer than 200 nt, often multi exonic and are engaged in a variety of functions like embryonic development (Ulitsky et al. 2011), dosage compensation (Chan et al. 2011), chromatin remodeling (Rinn et al. 2007) and may even act as scaffolds for chromatin remodelling complexes like PRC2 (involved in X-inactivation) (Jeon and Lee 2011) and influence gene regulation of miRNA target genes (Salmena et al. 2011). Comparative genomics may also yield insights into genomic rearrangements including events such as duplications, inversions, or translocations that change the structure of the genome. It has been hypothesized that genomic rearrangements could promote speciation by reducing gene flow by suppressing recombination and extending the effects of linked isolation genes (Rieseberg 2001).

Elephants are extraordinary mammals that have evolved numerous characteristics unique to them. Identifying the genetic bases of evolutionary changes is an extremely challenging task and comparative sequence analyses overcome many of these limitations but the availability of high quality genome and transcriptome sequence is critical to carry out such analyses. Despite the availability of the whole genome...
sequence of *L. africana*, most of the research on elephants in the past as indicated by published literature has been in the area of behaviour, ecology, reproductive biology and breeding, infectious disease, life history traits and welfare of wild and captive populations in addition to population genetics and demography, poaching, conflicts with people and other conservation issues.

At the time this project was conceived no whole genome sequences were available for the Asian elephant. Given the split between Asian and African elephants that occurred approximately 7 million years ago (Rohland et al. 2010) and their different evolutionary histories, we expected to find genetic differences between the two species. While *Loxodonta* remained confined to the African continent, *Elephas* migrated into Eurasia with a distinct phylogeography shaped by the glacial-interglacial cycles of the Pleistocene (Vidya et al. 2009). We therefore aimed to develop a resource for genome-wide sequence analyses of the Asian elephant and comparative analyses with the genome sequences of the African elephant, other members of Afrotheria as well as mammals in general. We carried out the whole genome and transcriptome sequencing of an Asian elephant from India. Earlier this year, high coverage sequences of three Asian elephants were reported (Lynch et al. 2015). However, the current study is the first report of high coverage transcriptome sequencing from the Asian elephant.

From the comparative analysis of genome sequences, we report a large number of Asian elephant specific, non-synonymous SNVs when compared to the African elephant and mammoth genomes. One of the most significant groups of genes in this dataset were the olfactory receptor (OR) genes. The African elephant genome has previously been shown to encode an expanded repertoire of OR genes when compared to other mammals (Niimura et al. 2014). The OR gene family is known to be rapidly evolving, with frequent gains (due to gene duplication) and losses (due to the conversion of functional copies into pseudogenes) (reviewed in Jiang and Matsunami 2015). The number of OR genes present in a genome has also been found to vary between different mammalian species, on the basis of adaptation to their respective environments (Hayden et al. 2010; Niimura 2012). For example, primates (including humans) have much fewer functional OR genes as well as pseudogenes, presumably because primates depend more on vision which is highly developed (Matsui et al. 2010). It is therefore not surprising to find species specific differences in more than a hundred OR genes between the Asian and the African elephant.

### Table 2.

<table>
<thead>
<tr>
<th>Cuff ID for ncRNA</th>
<th>Putative IncRNA</th>
<th>e-value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUFF.27.1</td>
<td>Pinc</td>
<td>0.01</td>
<td>Highly conserved mammalian ncRNA, involved in cell survival and cell cycle progression, discovered in mammary glands</td>
</tr>
<tr>
<td>CUFF.4121.1</td>
<td>bxd_6</td>
<td>0.00075</td>
<td>Involved in Regulation of Ultrabithorax via transcriptional interference</td>
</tr>
<tr>
<td>CUFF.5432.1</td>
<td>FMR1-AS1</td>
<td>0.0015</td>
<td>X linked FMR gene antisense RNA, reported to have anti apoptotic function</td>
</tr>
<tr>
<td>CUFF.7140.1</td>
<td>PVT1_3</td>
<td>0.0026</td>
<td>IncRNA located at translocation breakpoint in Burkitt’s lymphoma. PVT1 is an oncogene that generates a non-coding RNA</td>
</tr>
<tr>
<td>CUFF.10590.1</td>
<td>DLEU2_1</td>
<td>1.2E−24</td>
<td>Lymphocytic specific IncRNA, potential tumor suppressor gene implicated in acute B cell lymphocytic leukaemia</td>
</tr>
<tr>
<td>CUFF.13499.1</td>
<td>ZNRD1-AS1_1</td>
<td>0.0037</td>
<td>IncRNA encoded from the MHC-1 locus</td>
</tr>
<tr>
<td>CUFF.17087.1</td>
<td>Mico1</td>
<td>0.0051</td>
<td>Imprinted ncRNA involved in Circadian rhythms</td>
</tr>
<tr>
<td>CUFF.18583.1</td>
<td>MALAT1</td>
<td>6.9E−19</td>
<td>Highly conserved mammalian IncRNA ubiquitously expressed in all tissues. Regulates expression of metastasis related genes and cell motility</td>
</tr>
<tr>
<td>CUFF.18578.1</td>
<td>NEAT1_3</td>
<td>9.2E−35</td>
<td>Constitutively expressed in all tissues, structural component of nuclear paraspeckles. Also reported to be a virus induced cRNA (VINC)</td>
</tr>
<tr>
<td>CUFF.15862.1</td>
<td>KCNQ1OT1</td>
<td>0.0043</td>
<td>Paternally imprinted expression. The main function of this RNA is to maintain silencing of the paternally silenced KCNQ1 gene by recruiting PRC2 repressive complex to the enhancer in cis.</td>
</tr>
</tbody>
</table>

IncRNA names and description are from the RFam database. Cuff IDs are generated from the transcriptome assembly reported here.
Further investigation is necessary to determine whether the SNVs identified in this study result in functional differences in perception of olfactory cues between the two species. However, it is not difficult to imagine that as the ancestral *Elephas* migrated into the Asian land mass, they would have encountered significantly different environments, climate and vegetation patterns, glaciation events (Vidya *et al*. 2009) and other conditions that were different from those present in Africa. The differences in OR gene sequences between the Asian and African elephants...
therefore might possibly reflect a functional adaptation to the difference in various volatile chemical signals they encountered. For instance, the pheromone, (Z)-7-dodecen-1-yl acetate, released in urine by female Asian elephants to signal estrous status, is different from that of the African elephant (Rasmussen and Schulte 1998). In the future, it would also be interesting to closely examine the number and overlap of OR genes between the three elephantids (mammmoths, Asian and African), as well as between elephantids and other members of Afrotheria comprising of species that have distinct diets and occupy diverse habitats.

We explored genomic rearrangements in the *E. maximus* genome by looking at discordant paired end alignments on the loxAfr3 assembly of the paired ends reads from the *E. maximus* genome. The number of rearrangements is far greater in the smaller scaffolds (toward the 10–12 o’clock region in the figure), which indicates that the smaller scaffolds are sites where a better assembly than the current one may be obtained.

Since we had generated transcriptome data from a single individual at a single time point, we cross validated our data by comparing our WBC transcriptome with the fibroblast transcriptomes of the African elephant. As expected, our dataset showed a gene clustering pattern that was distinct from the two fibroblast transcriptomes, and an enrichment of immune system related genes. We then used our transcriptome sequence to perform transcriptome based annotation of the *L. africana* genome. Many genomes sequenced as a part of large sequencing projects carry out annotations using computational pipelines based on homology to human and mouse genes, that may not be able to identify species or lineage specific features. Based on our transcriptome data, we were able to identify more than a hundred novel expressed transcripts with protein coding potential and homology to known protein domains, most of them with homology to immune-system related genes or domains. These genes may either be present and/or expressed only in the Asian elephant, or they may be found in the African elephant as well, but were not annotated previously. Validation of these novel transcripts would reveal whether elephants (in general, or Asian elephants in particular) might have evolved novel immune system related functions. Additional studies on comparative analyses of tissue specific and developmental stage specific transcriptomes would further improve the annotation of the elephant genomes and help identify many more species specific genes, as well as gene expression patterns. A large number of transcripts that were expressed could not be identified or annotated based on sequence and protein domain comparisons. One possibility is that such transcripts could represent non-coding RNAs which do not contain protein coding potential.

We therefore investigated the possibility of expression of non-coding RNAs, and particularly, IncRNAs in the transcriptome. Our study is the first to report the expression of IncRNAs in the Asian elephant. Evidence that IncRNAs perform biologically relevant roles in organisms has accrued from studies that demonstrate the dysregulation of IncRNAs in disease conditions (Gibb et al. 2011; Mitra et al. 2012). Mutations in the IncRNA sequence and alterations to DNA methylation at IncRNA promoters have been demonstrated to be linked to cancer (Du et al. 2013; Zhi et al. 2014). Taken together all these facts about IncRNAs paint a compelling picture of the very important yet largely unexplored depths of what IncRNAs are involved in. The occurrence of highly conserved IncRNAs in the Asian elephant transcriptome provides confidence in the quality of our dataset in representing a comprehensive transcriptome of the PBMCs in *E. maximus*. These IncRNAs need to be functionally validated in an in vivo scenario by functional screens. cDNA cloning of these sequences into mammalian expression vectors and then transfection into other mammalian cells might be one way of performing such analysis.

Significant number of IncRNAs could not be identified with the RFam database searches and the reason for this could be two-fold. The RFam database may not be comprehensive enough to harbor entries for all possible IncRNAs for reasons explained above (see sections 2.8 and 3.5). Secondly, some of the transcripts that have been tagged as putative IncRNAs may be novel and thus have not been reported earlier. An in-depth analysis with functional validation is the only way in which this could be resolved.

Comparative analyses of domain architectures between the available proteomes of *L. africana* with three other Afrotherian species showed an increased number of proteins in almost every domain category in the African elephant, despite its evolutionary proximity to these species. As expected, we observed the African elephant protein domain architectures overlapped most with those of the manatee, which is evolutionarily closer to elephants than other Afrotheria species. We also identified 181 proteins with domain architectures that were unique to the African elephant. Further analyses of these proteins and their functions might yield interesting insights into elephant specific traits. It would also be interesting to perform a similar comparison between the different elephantids when the assembly and proteome of the Asian elephant and woolly mammoth become available.

These findings have broader implications in furthering the understanding of a wide range of genetic and evolutionary processes among elephants in particular, and Afrotheria and mammals in general. Our data serves as a primary resource that can be analysed by researchers from various fields to extract specific information that would be of relevance to answering questions of their interest. Further analyses and comparison of whole genome and transcriptome sequences would help unravel the evolutionary history of many more unique traits that may have conferred an adaptive advantage.
to the elephantid lineage and enhance our understanding of how these magnificent creatures evolved and diverged. For example, studies focused on key genes of the immune system can be carried out to understand the susceptibility or adaptations of Asian elephants to various types of pathogens, irritants, allergens and toxins. Genetic predispositions to diseases and physiological susceptibility to various environmental and anthropogenic factors can also be studied. This report also provides a stepping-stone for evolutionary biologists to study the evolution of elephant specific traits such as large body size, their metabolic machinery and skeletal structure, tusks and aspects related to olfaction.

Acknowledgements

This work was supported by grant from the Centre of Excellence in Epigenetics program of the Department of Biotechnology, India, and from IIISER Pune to SG. PCR is supported by Postdoctoral Fellowship from the Department of Biotechnology, India. AK is supported by an Early Career Fellowship from the Wellcome Trust-DBT India Alliance. FH is supported by the Centre of Excellence in Epigenetics Program. SJP is supported by Senior Research Fellowship from the Council of Scientific and Industrial Research, India. IS is supported by Centre for Ecological Sciences through grant-in-aid from Ministry of Environment, Forest and Climate Change. RS is a JC Bose National Fellow. Elephant blood sample was obtained with the permission and support of the Karnataka Forest Department (vide permit No. B1/WL/CR-25/2013-14).

References

Asher RJ and Lehmann T 2008 Dental eruption in afrotherian mammals. BMC Biol. 6 14
Carter AM, Blankenship TN, Enders AC and Vogel P 2006 The fetal membranes of the otter shrews and a synapomorphy for afrotheria. Placenta 27 258–268

da Huang W, Sherman BT and Lempicki RA 2009 Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 4 44–57
Dastjerdi A, Robert C and Watson M 2014 Low coverage sequencing of two Asian elephant Elephas maximus genomes. Gigascience 3 12
Huang DW, Sherman BT and Lempicki RA 2009 Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37 1–13
Jeon Y and Lee JT 2011 YY1 tethers Xist RNA to the inactive X chromosome. Cell 146 119–133

J. Biosci. 40(5), December 2015


Oliveros J 2007 VENN¥ An interactive tool for comparing lists with Venn diagrams.


Quinlan AR and Hall IM 2010 BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26 841–842

R Core Team 2014 R: A language and environment for statistical computing R Foundation for Statistical Computing, Vienna, Austria (http://www.R-project.org)


J. Biosci. 40(5), December 2015
Novel transcripts and variants in the Asian elephant


Sulak M, Fong L, Mika K, Chigurupati S, Yon L, Mongan NP et al. 2015 TP53 copy number expansion correlates with the evolution of increased body size and an enhanced DNA damage response in elephants bioRxiv. doi:10.1101/028522


Svartman M and Stanyon R 2012 The chromosomes of Afrotheria and their bearing on mammalian genome evolution. Cytogenet Genome Res. 137 144–153


MS received 03 November 2015; accepted 16 November 2015

Corresponding editor: PARTHA P MAJUMDER

J. Biosci. 40(5), December 2015