Table 1 | *C. atys* assembly statistics and proteins with major structural variations in the *C. atys* genome

Assembly		Annotation	
Average coverage per base	192	Protein-coding genes	20,829
Total sequence length	2,848,246,356 bp	Non-coding genes	4,464
Total assembly gap length	60,973,502 bp	Pseudogenes	5,263
Number of scaffolds	11,433	mRNA transcripts	65,920
Scaffold N50	12,849,131 bp	lncRNA transcripts	6,299
Scaffold L50	66	Exons in coding transcripts	250,660
Number of contigs	76,752	Exons in non-coding transcripts	42,280
Contig N50	112,942 bp		
Contig L50	6,930		
GC content	40.90%		

Gene	Function	Variation type	Length variation (amino acids)
ICAM2	Lymphocyte extravasation and recirculation	indel, fs	107
TLR4	LPS sensing	indel, fs	17
BPIFA1	Antimicrobial function in airways	indel	8
NOS2	Proinflammatory messenger	pm, early stop	8
MBL2	Pattern recognition receptor for microbial products	pm, early start	7
TREM2	Chronic proinflammatory signalling in myeloid cells	indel, fs	6
PLSCR1	Enhancement of the interferon response	indel	5
LST1	Inhibition of lymphocyte proliferation	indel, fs	5
CRTAM	T and natural killer cell activation	pm, indel	4

Structural variations were identified by the immunogenomic comparison pipeline. N50, 50% of the genome is in fragments of this length or longer; L50, smallest number of fragments needed to cover more than 50% of the genome; lncRNA, long non-coding RNA; indel, insertion/deletion; fs, frameshift; pm, point mutation.

threshold of identity (Extended Data Fig. 1b, c). In addition, we found specific gene families in *C. atys* that are expanded relative to *M. mulatta*, humans and other primates (Extended Data Table 2a). Notably, we detected localized regions of increased substitution, defined by a clustered difference of three or more amino acids, in 10 genes. The most marked variations in the amino acid sequence of *C. atys* compared to *M. mulatta* were observed in ICAM-2 and TLR-4 (Table 1).

ICAM-2 is an approximately 60-kDa transmembrane glycoprotein of the immunoglobulin superfamily, which is expressed on various immune cells and implicated in lymphocyte homing and recirculation[6]. ICAM-2 ligands are lymphocyte function-associated antigen-1 and the C-type lectin DC-SIGN[7]. We discovered a misalignment of the ICAM-2 proteins between *C. atys* and *M. mulatta* that starts in exon 3 (Extended Data Fig. 2a). This difference is explained by a 499-bp deletion starting from exon 3 of Ca*ICAM2*, as detected by PCR and Sanger sequencing (Fig. 2a and Extended Data Fig. 3). We subsequently confirmed the expression of this truncated form of ICAM-2 in ten out of ten additional *C. atys* genome sequences (Extended Data Fig. 2b). By contrast, analysis of the whole-genome sequences of 15 baboons and more than 130 rhesus macaques demonstrated that only the full-length ICAM-2 protein was found in all individuals (data not shown)[8]. The ICAM-2 deletion may be specific to *C. atys*, as it is not present in any other known primate sequences, including other natural SIV hosts, such as the African green monkey, drill and colobus monkey. Transcript models generated from *de novo* assembled *C. atys* RNA-sequencing (RNA-seq) data from 14 different tissues showed that the mature mRNA sequence of Ca*ICAM2* retains substantial portions of what is part of the intronic sequence in other nonhuman primates, and thus codes for a markedly different final gene product (Extended Data Figs 2, 3). Splice-junction sequence analysis showed intact splicing for all four exons in *M. mulatta*, but no splice junctions were found between exons 3 and 4 in *C. atys*, indicating severe splicing defects due to the deletion (Extended Data Fig. 4).

To test whether the observed genetic difference in *ICAM2* has functional consequences, we measured ICAM-2 surface expression on immune cells from humans, *M. mulatta* and *C. atys* with an antibody that recognizes a conserved epitope between these species[9]. ICAM-2 was readily detected on T cells and B cells from humans and *M. mulatta*, but not from *C. atys* (Fig. 2b, c), suggesting that ICAM-2 is not functional in lymphocytes of *C. atys*. However, a truncated, lower

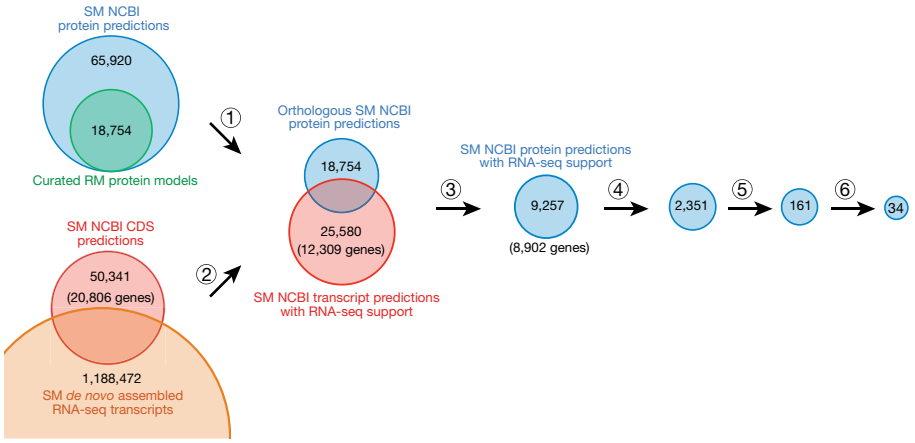

Figure 1 | Bioinformatic pipeline for the identification of divergent *C. atys* proteins. (1) Sooty mangabey (SM) orthologues were selected by BLAST alignment of *C. atys* NCBI protein predictions (blue) to curated rhesus macaque (RM) protein models (green[22]) and alignment scores were calculated. (2) NCBI transcript predictions with RNA-seq support were identified by BLAT alignment of *de novo* assembled *C. atys* RNA-seq transcripts (orange) to *C. atys* NCBI coding sequence (CDS) predictions (red). (3) Subsquently, corresponding RNA-seq-supported *C. atys* NCBI protein predictions were selected. (4) *C. atys* proteins with high similarity (>97% identity) to *M. mulatta* proteins were filtered out. (5) Immune genes according to Gene Ontology (GO) term classification (immune response) were chosen for further analysis and (6) confirmed by manual inspection.