PNAS
www.pnas.org

Supplementary Information for

Human Exonization through Differential Nucleosome Occupancy

Yumei Li[*], Chen Li[*], Shuxian Li, Qi Peng, Ni A. An, Aibin He[#], and Chuan-Yun Li[#]

Chuan-Yun Li, Email: chuanyunli@pku.edu.cn
Aibin He, Email: ahe@pku.edu.cn

This PDF file includes:

Supplementary text
Figs. S1 to S9
Tables S1 to S3
References for SI reference citations

Other supplementary materials for this manuscript include the following:

Datasets S1 to S2

www.pnas.org/cgi/doi/10.1073/pnas.1802561115

Supplementary Materials and Methods

Sample collection. Tissue samples used in this study were obtained from the animal facility of the Institute of Molecular Medicine in Peking University (accredited by the Association for Assessment and Accreditation of Laboratory Animal Care). The samples were obtained in accordance with protocols approved by the Animal Care and Use Committee of Peking University (IMM-HeAiB-1).

Chromatin digestion by MNase, library preparation, and deep sequencing. MNase-seq was performed as previously described (1). Briefly, fresh tissues were immediately fixed in 1% v/v formaldehyde at room temperature for 10 min, neutralized with 2.5 M glycine (Sigma), and washed with PBS. Then the fixed tissues were homogenized with a T10 homogenizer (IKA) and collected by centrifugation. The collected samples were incubated in hypotonic buffer for 20 min and homogenized by stirring 10-20 strokes. Pellets were washed with MNase digestion buffer (in mM: 10 Tris pH 7.4, 15 NaCl, 10 KCl, 1.5 $CaCl_2$) and centrifuged at 350×g for 5 min at 4 °C. We then suspended the collected pellets in digestion buffer containing MNase (New England Biolabs), incubated them at 37 °C, and stopped the reaction with 2× Stop Buffer (0.6% SDS, 40 mM EDTA, 40 mM EGTA). We determined the concentration of MNase and the digestion time for each sample to ensure that about 80% mononucleosomes were present after digestion (**SI Appendix, Table S3**). The products of MNase digestion were then reverse crosslinked at 65 °C overnight and treated with RNaseA (Sigma) at 37 °C for 30 min, followed by a treatment with proteinase K at 55 °C for 6 h. DNA was then extracted using the phenol-chloroform method, with DNA fragments ranging from 100 to 200 bp collected and purified with a Qiagen Gel Extraction Kit. The MNase-seq libraries were then prepared using a NEBNext DNA Sample Preparation Kit (New England Biolabs) and the samples were sequenced on an Illumina Hiseq sequencing system in the paired-end mode.

RNA extraction, library preparation, and deep sequencing. Total RNAs from the same tissue samples as those from the MNase-seq were extracted using the Trizol method and analyzed on an Agilent 2100 Bioanalyzer (Agilent Technologies). Strand-specific, poly (A)-positive RNA-seq libraries were then prepared and sequenced with a paired-end design on the Illumina HiSeq platform, according to the manufacturer's instructions.

Sequencing data analysis. MNase-seq reads were filtered with in-house scripts to obtain and retain high-quality reads that satisfied the following criteria: fraction of N < 5%; average quality of reads > 20; fraction of low-quality bases (Phred score < 20) in a read < 50%. The filtered reads were then aligned to the corresponding genomes (hg19, rheMac2, tupBel1, mm9 and susScr3 for human, rhesus macaque, tree shrew, mouse and mini pig, respectively) using BWA (version 0.7.13-r1126) (2). The uniquely mapped reads with mapping quality ≥ 30 and fragment size between 100 and 250 bp were retained for further analyses. DANPOS2 (version 2.2.2) was then used to call nucleosome binding peaks and to generate nucleosome occupancy profiles (-jd 147 -m 1 --extend 74) (3), which were further normalized with the mean score of the whole genome.

We then introduced several previously-described parameters to evaluate nucleosome occupancy profiles (1, 4-6). These included fragment size distribution of the fragment sizes of all the uniquely mapped, paired-end reads; dinucleotide frequency as estimated by the AA/AT/TA/TT and CC/CG/GC/GG dinucleotide frequency across all 147 bp fragments; and the nucleosome phase across gene TSS, in which the nucleosome occupancy of chromosome regions that are adjacent to gene transcription start sites (±1,500 bp) were calculated, averaged, and compared.

Strand-specific, poly (A)-positive RNA-seq data were analyzed as previously described (7). Briefly, raw reads were filtered using the same criteria as that used for the MNase-seq data and then they were aligned to the corresponding genome using TopHat2 (TopHat v2.1.1) (8). Only uniquely mapped reads were used in the following analyses.

Whole-genome comparison of nucleosome occupancy profiles. To obtain a global view of the nucleosome occupancy profiles across both tissues and species, we introduced a method to evaluate the similarity of the nucleosome distribution. Briefly, we obtained the orthologous regions across all five species using the multiple alignment results from the UCSC Genome Browser (9). These orthologous regions were then divided into 147-bp windows. For each dataset of nucleosome profiling, the summit positions of the nucleosome binding peaks were recorded for each 147-bp window, and the distances between the peak summit and the starting position of the peak-located window were calculated. The distance was designated “NA” if no nucleosome binding signals were located in the window. A distance vector for each nucleosome occupancy profile was then obtained. We next calculated the Pearson correlation coefficient between the distance vectors for each pair of profiles. Hierarchical clustering was used to evaluate the similarity of these nucleosome occupancy profiles.

Identification of recently-evolved human exons. The initial list of human exons was compiled on the basis of annotations from multiple databases (such as RefSeq (10), Ensembl (11), UCSC (12), Genescan (13), and Vega (14)), in which only internal exons were considered. These exons were then filtered using RNA-seq data for human brain to only retain exons supported by junction reads on both sides of the exon (7). For each human exon, the orthologous regions in macaque and mouse were extracted based on pair-wise whole genome alignments, retaining exons uniquely aligned as a fragment in out-group species, a strict strategy to control for false-positives obtained by gene duplication. Candidate recently-evolved human exons were then defined on the basis of annotations and RNA-seq data, in which the orthologous regions of candidate recently-evolved human exons in out-group species could not overlap with any annotated exons (annotations from RefSeq, Ensembl and RhesusBase (15, 16) for macaque; annotations from RefSeq, Ensembl, UCSC, Genescan, and Vega for mouse) (12), and was not supported by any in-house RNA-seq data in tissues of macaque and mouse (no supporting junction reads and an RPKM score < 0.2) (17).

To further control for the potential false-positives, we performed ultra-deep, strand-specific, poly (A)-positive RNA-seq on the macaque brain sample to increase the detection sensitivity of macaque exons. A total of 1.4 billion RNA-seq reads were generated and uniquely mapped to the macaque genome using a computational pipeline as previously described (7). We also integrated more public RNA-seq data of rhesus macaque and mouse – a total of 97 RNA-seq datasets from 18 types of macaque tissues, as well as 64 RNA-seq datasets from 13 types of mouse tissues – to control for the false-positives introduced by the variability among populations and tissue types (15). Candidate recently-evolved human exons supported by any of these RNA-seq datasets in out-group species were removed.

NOC ratio and GC content ratio. We introduced NOC ratio and GC content ratio to quantitatively investigate exon-intron differences in nucleosome occupancy and GC content, respectively. The NOC ratio was calculated by binary logarithm of the average nucleosome occupancy of exon to that of the upstream 150-bp intronic region. Similarly, the GC content ratio was calculated by binary logarithm of the GC content of exon to that of the upstream 150-bp intronic region.

Ancestral state definition and the calculation of interspecies divergence rate. When counting the divergence sites in human and macaque lineages after the divergence of the two species, multiple sequence alignment data were downloaded from the UCSC Genome Browser (12), and the human-macaque-marmoset aligned regions were extracted to define the ancestral state of human and macaque at each site as previous described (18, 19). For simplicity, we used AT-to-GC to stand for A:T to G:C and A:T to C:G substitutions, while GC-to-AT for G:C to A:T or G:C to T:A substitutions.

As for the inference of the ancestral state for nucleotide sites nearby the splice sites, orthologous sequences from 46 vertebrate genomes species downloaded from the UCSC Genome Browser were used to infer the common ancestor of human, macaque, and mouse on the basis of the FASTML pipeline (20). Positions with indefinable ancestral sequences were excluded. The strengths of the splice donor and acceptor sites were then calculated using maxEndScan (21), and the sequence logo of splice sites were visualized by WebLogo (22).

Code availability

All the code used to generate results can be found at GitHub via URL https://github.com/xihuimeijing/Nucleosome-and-exon-evolution.

Author contributions.

CYL, AH, and YL conceived the idea and designed the study. YL and CL analyzed data and performed most of the analyses. SL, QP, and NA performed part of the analyses. CYL and YL wrote the paper. All authors read and approved the final manuscript. The authors declare no competing financial interests.

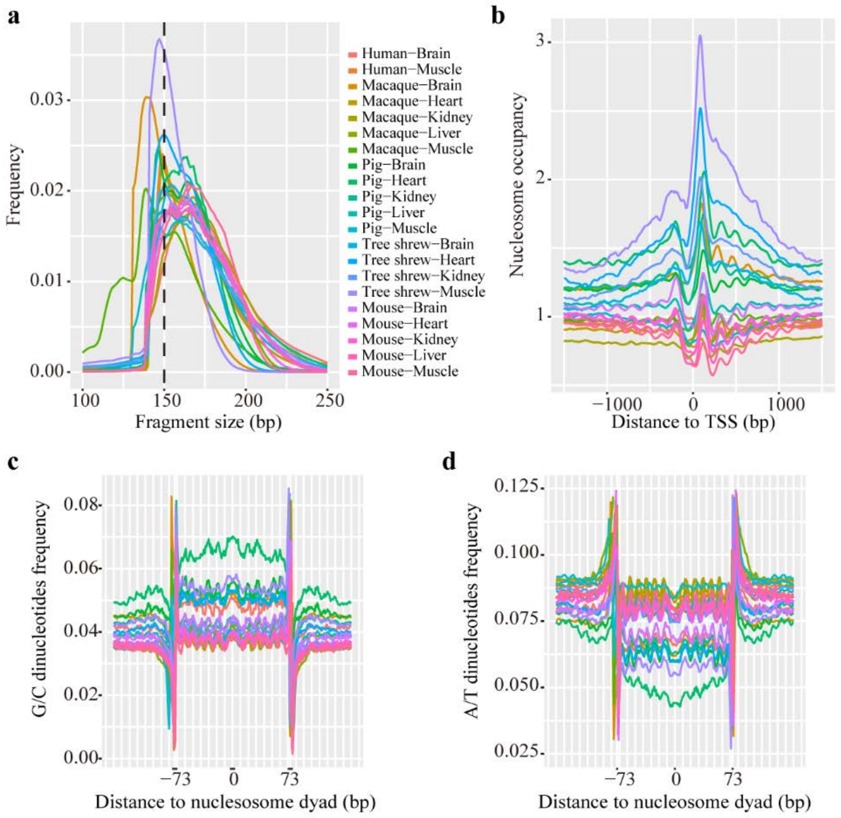

Fig. S1. High-quality nucleosome occupancy profiles for mammalian tissues. (a) Plot for the fragment size distribution of MNase-seq, with 150 bp indicated with dotted line. **(b)** Aggregate lines for nucleosome occupancy profiles near the gene transcription start sites (TSS) are shown. **(c-d)** For nucleosome-protected regions in all tissue samples of the five species, the dinucleotide frequency of G/C (the combined dinucleotide frequency of CC/CG/GC/GG) **(c)** and A/T (the combined dinucleotide frequency of AA/AT/TA/TT) **(d)** are shown. Nucleosome dyad, the midpoint position of the DNA bound by the nucleosome core.

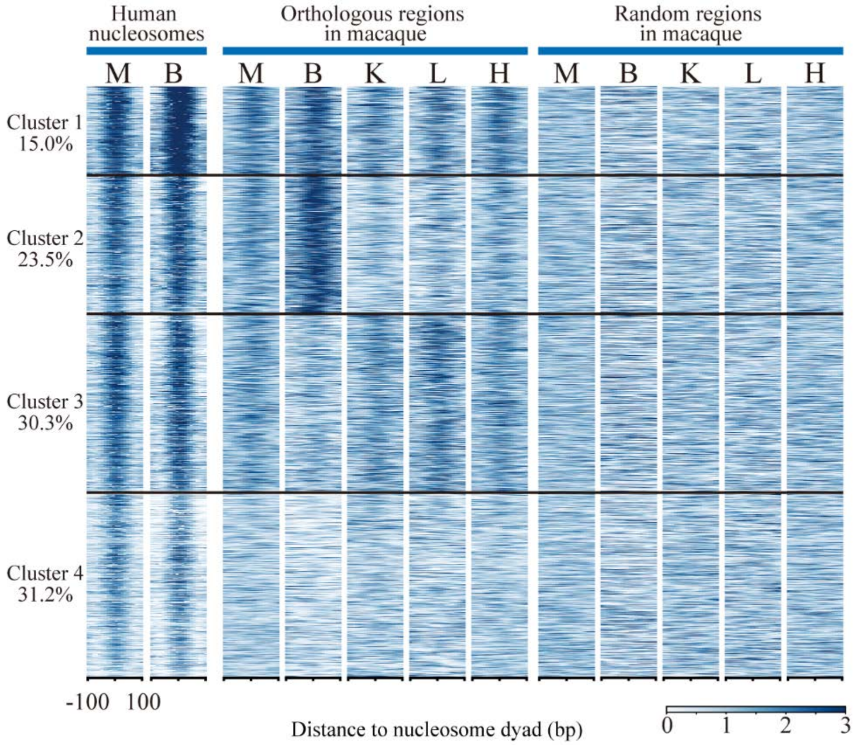

Fig. S2. Clustered heat map of nucleosome occupancy profiles. Heat map of nucleosome occupancy profiles for human samples and nucleosome occupancy profiles at macaque orthologous regions, as well as at properly aligned, randomly selected macaque genomic regions. The heat maps were clustered by k-means method, with the proportion of each cluster indicated. **B**, brain; **H**, heart; **K**, kidney; **L**, liver; **M**, muscle.

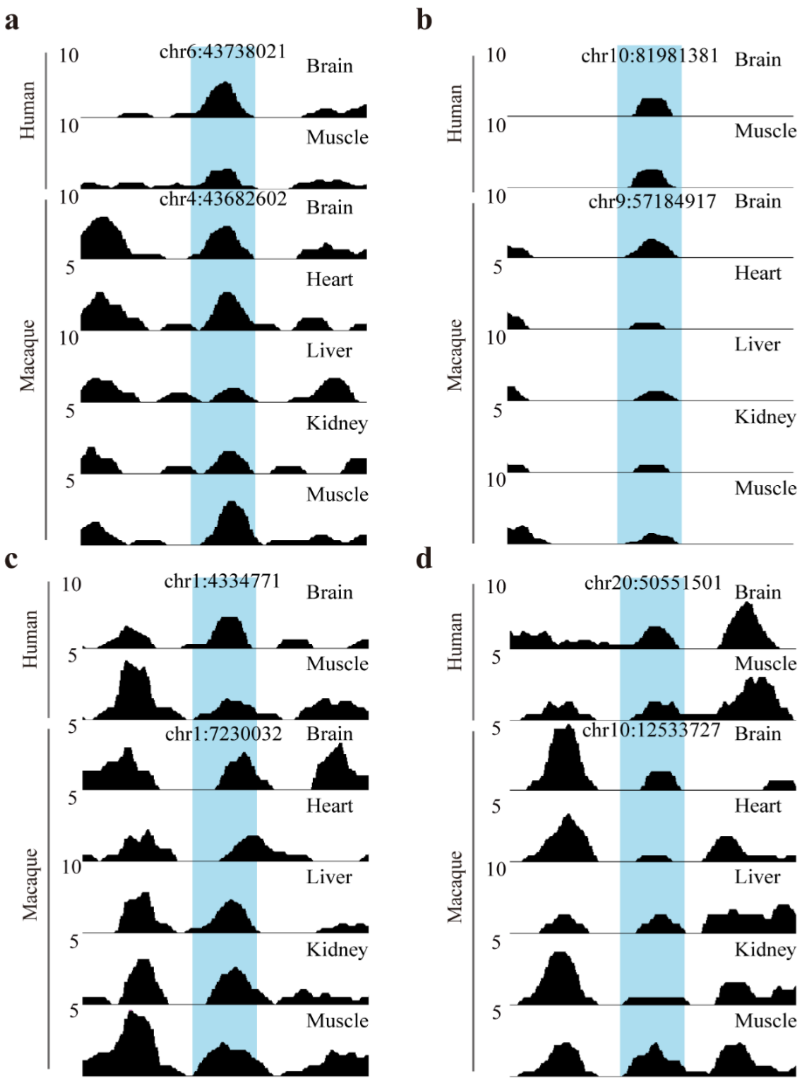

Fig. S3. Demonstration cases for cross-tissue and cross-species conservation of the nucleosome occupancy profiles. Nucleosome occupancy profiles of two human tissues in four genomic regions together with the profiles of multiple macaque tissues in the orthologous regions are shown. In each profile, a 147-bp window, shaded in blue, indicates a region occupied by one nucleosome with the coordinate of its midpoint shown above each panel.

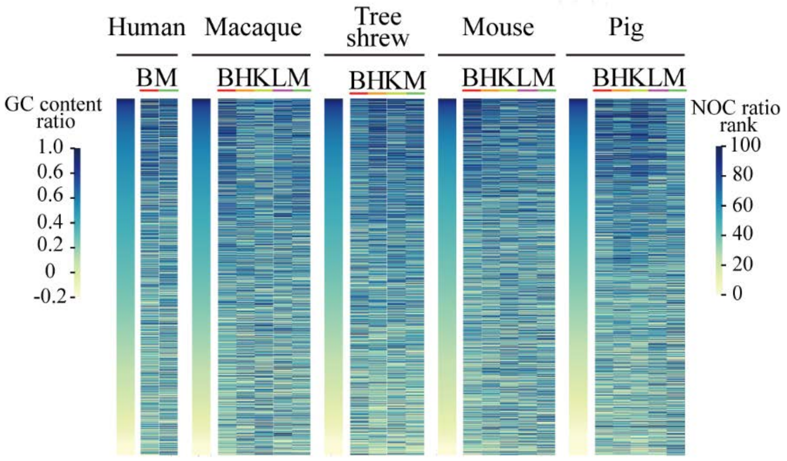

Fig. S4. Heat maps of the exon-intron differences in GC content and nucleosome occupancy in different species. The exon-intron differences in GC content and nucleosome occupancy were measured by GC content ratios and NOC ratios, respectively. For each species, the map was sorted by decreasing order of GC content ratio (**left panel**), the NOC ratios for the corresponding regions were then shown accordingly (**right panels**). **B**, brain; **H**, heart; **K**, kidney; **L**, liver; **M**, muscle.

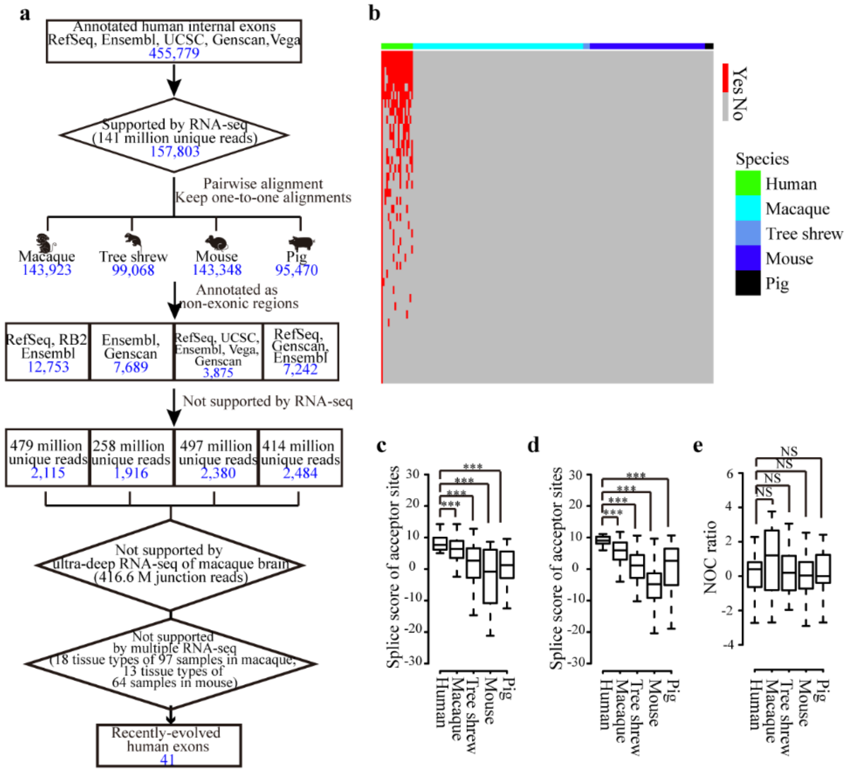

Fig. S5. Differential nucleosome occupancy appears prior to the origination of recently-evolved human exons. **(a)** Flowchart for the identification of recently-evolved human exons on the basis of RNA-seq data and gene annotations. The number of human exons in each step of the flowchart is indicated in blue. **RefSeq**, gene annotations from NCBI Reference Sequence Database (10); **Ensembl**, gene annotations from Ensembl (11); **UCSC**, gene annotations from UCSC genome browser (12); **Genscan**, gene predictions from Genscan program (13); **Vega**, gene annotations from the Vertebrate Genome Annotation database (14); **RB2**, gene annotations from RhesusBase (version 2) (15). **(b)** Heat map showing the expression of the candidate recently-evolved human exons in different human individuals, as well as the expression of the orthologous regions in macaque, tree shrew, mouse and pig animals. Horizontal red bars indicate the existence of exon expression. **(c-e)** Splice site scores of acceptor sites **(c)**, donor sites **(d)**, and NOC ratios **(e)** are shown in boxplots, for recently-evolved human exons in human and orthologous regions in macaque, tree shrew, mouse and pig. NS, not significant; *P ≤0.05; **P ≤0.01; ***P ≤0.001.

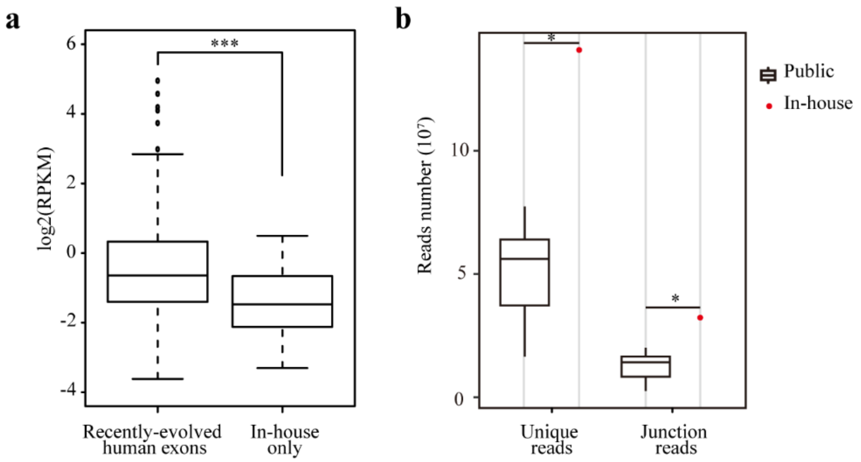

Fig. S6. Comparisons of public and in-house RNA-seq data. **(a)** Boxplots showing the expression levels of recently-evolved human exons (**Recently-evolved human exons**), as well as a sub-list of these exons detected in in-house RNA-seq data but not in public datasets (**In-house only**). **(b)** Boxplots showing the numbers of the uniquely-mapped reads (**Unique reads**) and junction reads (**Junction reads**) of in-house (**In-house**) and public RNA-seq datasets (**Public**). NS, not significant; **P* ≤0.05; ***P* ≤0.01; ****P* ≤0.001.

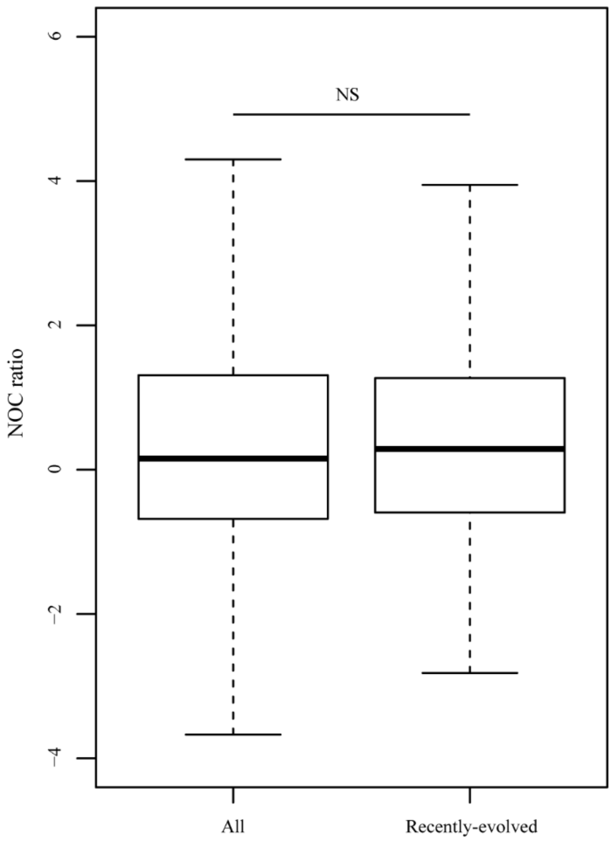

Fig. S7. NOC ratios of all annotated and recently-evolved human exons. NOC ratios are shown in boxplot for all annotated human exons (**All**) as well as recently-evolved human exons (**Recently-evolved**). NS, not significant.

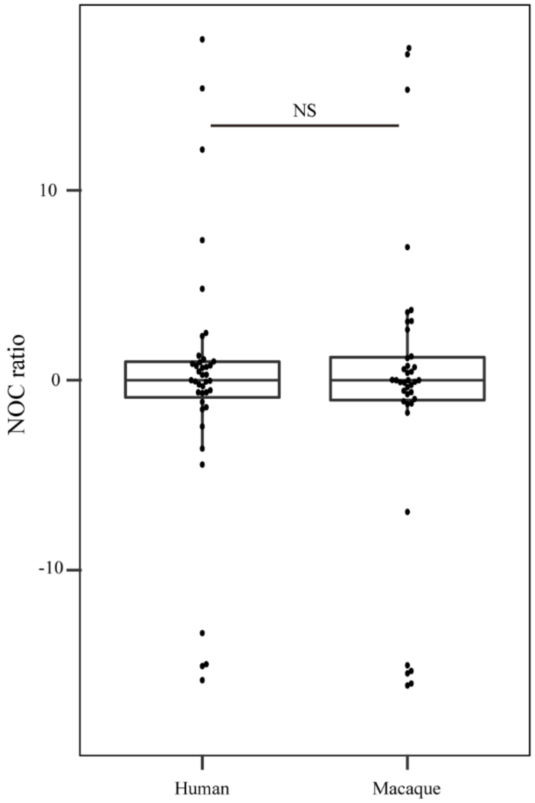

Fig. S8. NOC ratios of recently-evolved human exons in human and rhesus macaque. The distributions of NOC ratios of recently-evolved human exons are shown. For these exons, the AG-GY boundaries could not be found in the orthologous regions of rhesus macaque, crab-eating macaque, baboon, squirrel monkey, tarsier, mouse lemur, bush baby, and mouse. NS, not significant.

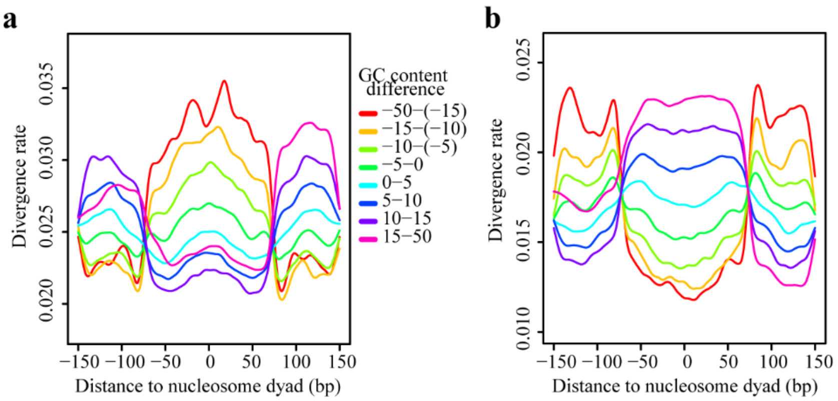

Fig. S9. Different divergence rates for nucleosome-protected regions with different GC content. Divergence rate of GC-to-AT **(a)** and AT-to-GC **(b)** substitutions were shown for nucleosome-protected regions in human after the divergence of human and rhesus macaque. All the nucleosome-protected regions were classified into eight categories based on the difference of GC content between these regions and their flanking linker regions. Nucleosome dyad, the midpoint position of the DNA bound by the nucleosome core.

Table S1. Summary of MNase-seq and RNA-seq in multiple tissues of the five species.

Species	Tissue	MNase-seq		RNA-seq	
		Total Reads (M)	Uniquely Mapped Reads (M)	Total Reads (M)	Uniquely Mapped Reads (M)
Human	Brain	203.4	194.2	159.3	140.9
	Muscle	300.3	289.2	143.2	122.2
Macaque	Brain	241.7	200.0	132.4	109.6
				1874.8	**1372.6**
	Heart	280.3	266.1	120.5	92.5
	Kidney	236.8	224.1	109.2	88.8
	Liver	178.8	168.6	116.5	91.8
	Muscle	367.6	342.9	122.3	96.4
Tree shrew	Brain	205.7	171.3	136.7	70.3
	Heart	231.9	188.9	145.7	58.1
	Kidney	219.8	179.4	139.2	69.8
	Muscle	102.8	76.8	135.7	60.2
Mouse	Brain	442.6	429.3	128.8	107.5
	Heart	364.8	351.4	143.9	105.5
	Kidney	230.6	222.7	132.7	106.2
	Liver	450.7	437.2	116.6	90.3
	Muscle	449.8	440.5	116.7	87.1
Pig	Brain	283.2	242.0	115.3	89.4
	Heart	327.4	246.7	81.0	64.1
	Kidney	307.1	262.1	93.9	72.6
	Liver	223.9	195.7	95.5	70.7
	Muscle	190.8	172.8	164.1	116.8

Table S2. Summary of public RNA-seq data used for the validation of recently-evolved human exons. Human RNA-seq data were downloaded from the GTEx project (23). RNA-seq data from rhesus macaque and mouse were extracted from RhesusBase (15, 16).

Species	Tissue	Sample Number
Human	Brain	18
Macaque	Brain	9
Macaque	Liver	13
Macaque	Kidney	7
Macaque	Heart	7
Macaque	Muscle	5
Macaque	Colon	3
Macaque	Lung	4
Macaque	Spleen	4
Macaque	Fat	1
Macaque	Testis	8
Macaque	Cerebellum	4
Macaque	Thymus	1
Macaque	Hippocampus	4
Macaque	Frontal pole	4
Macaque	Lymphoblastoid	9
Macaque	Caudate nucleus	6
Macaque	Frontal cerebral cortex	7
Macaque	Orbital cerebral cortex	1
Mouse	Brain	8
Mouse	Liver	7
Mouse	Kidney	7
Mouse	Heart	6
Mouse	Muscle	4
Mouse	Colon	3
Mouse	Lung	3
Mouse	Spleen	3
Mouse	Testis	5
Mouse	Cortical plate	5
Mouse	Subventricular zone	5
Mouse	Ventricular zone	5
Mouse	Cerebellum	3

Table S3. The experimental conditions for chromatin digestion by MNase.

Species	Tissue	Fixed time (min)	Started volume (μl)	Digestion system (μl)	MNase concentration (U/μl)	MNase digestion time (min)
Human	**Brain**	10	20	250	0.1	22
	Muscle	10	40	400	0.1	36
Macaque	**Kidney**	10	50	500	0.5	22
	Brain	15	25	250	0.1	20
	Heart	10	50	250	0.548	20
	Muscle	15	40	400	0.1	36
	Liver	10	30	400	0.5	20
Tree shrew	**Kidney**	10	50	500	0.5	22
	Brain	10	30	250	0.1	22
	Heart	10	50	250	0.548	18
	Muscle	10	30	250	0.1	36
Mouse	**Kidney**	10	150	1000	0.5	20
	Brain	10	250	1000	0.155	20
	Heart	10	200	1000	0.548	14
	Muscle	10	250	1000	0.065	20
	Liver	10	150	1000	0.5	18
Pig	**Kidney**	10	50	500	0.5	20
	Brain	10	30	250	0.1	20
	Heart	10	40	250	0.548	20
	Muscle	10	25	400	0.1	36
	Liver	10	50	500	0.5	18

Dataset S1. Coordinates and gene annotations for recently-evolved human exons using macaque and mouse as out-group species. The coordinates are based on genome version hg19 (GRCh37).

Dataset S2. Coordinates and gene annotations for recently-evolved human exons using macaque, tree shrew, mouse, and pig as out-group species. The coordinates are based on genome version hg19 (GRCh37).

References

1. Cui K & Zhao K (2012) Genome-wide approaches to determining nucleosome occupancy in metazoans using MNase-Seq. *Methods Mol Biol* 833:413-419.
2. Li H & Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754-1760.
3. Chen KF, *et al.* (2013) DANPOS: Dynamic analysis of nucleosome position and occupancy by sequencing. *Genome research* 23(2):341-351.
4. Satchwell SC, Drew HR, & Travers AA (1986) Sequence Periodicities in Chicken Nucleosome Core DNA. *Journal of molecular biology* 191(4):659-675.
5. Schones DE, *et al.* (2008) Dynamic regulation of nucleosome positioning in the human genome. *Cell* 132(5):887-898.
6. Brogaard K, Xi L, Wang JP, & Widom J (2012) A map of nucleosome positions in yeast at base-pair resolution. *Nature* 486(7404):496-501.
7. Zhang SJ*, et al.* (2017) Isoform Evolution in Primates through Independent Combination of Alternative RNA Processing Events. *Molecular biology and evolution* 34(10):2453-2468.
8. Kim D*, et al.* (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* 14(4).
9. Thomas JW*, et al.* (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* 424(6950):788-793.
10. Pruitt KD, Tatusova T, & Maglott DR (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research* 33(Database issue):D501-504.
11. Hubbard T*, et al.* (2002) The Ensembl genome database project. *Nucleic acids research* 30(1):38-41.
12. Karolchik D*, et al.* (2003) The UCSC Genome Browser Database. *Nucleic acids research* 31(1):51-54.
13. Burge C & Karlin S (1997) Prediction of complete gene structures in human genomic DNA. *Journal of molecular biology* 268(1):78-94.
14. Ashurst JL*, et al.* (2005) The Vertebrate Genome Annotation (Vega) database. *Nucleic acids research* 33(Database issue):D459-465.
15. Zhang SJ*, et al.* (2014) Evolutionary interrogation of human biology in well-annotated genomic framework of rhesus macaque. *Molecular biology and evolution* 31(5):1309-1324.

16. Zhang SJ*, et al.* (2013) RhesusBase: a knowledgebase for the monkey research community. *Nucleic acids research* 41(Database issue):D892-905.

17. Chen JY*, et al.* (2015) Emergence, Retention and Selection: A Trilogy of Origination for Functional De Novo Proteins from Ancestral LncRNAs in Primates. *Plos Genetics* 11(7).

18. Chen X*, et al.* (2012) Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes. *Science* 335(6073):1235-1238.

19. Prendergast JG & Semple CA (2011) Widespread signatures of recent selection linked to nucleosome positioning in the human lineage. *Genome research* 21(11):1777-1787.

20. Ashkenazy H*, et al.* (2012) FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic acids research* 40(W1):W580-W584.

21. Yeo G & Burge CB (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of Computational Biology* 11(2-3):377-394.

22. Crooks GE, Hon G, Chandonia JM, & Brenner SE (2004) WebLogo: A sequence logo generator. *Genome research* 14(6):1188-1190.

23. Lonsdale J*, et al.* (2013) The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* 45(6):580-585.