Positive selection of the TRIM family regulatory region in primate genomes

Dan-Dan He1,†, Yueer Lu1,†, Rachel Gittelman2, Yabin Jin1, Fei Ling1 and Akey Joshua2

1 School of Bioscience and Bioengineering, South China University of Technology, Guangzhou 510006, People’s Republic of China
2 Department of Genome Sciences, University of Washington, Seattle, WA 98125, USA

† These authors contributed equally to this study.

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.3491580.

Viral selection pressure has acted on restriction factors that play an important role in the innate immune system by inhibiting the replication of viruses during primate evolution. Tripartite motif-containing (TRIM) family members are some of these restriction factors. It is becoming increasingly clear that gene expression differences, rather than protein-coding regions changes, could play a vital role in the anti-retroviral immune mechanism. Increasingly, recent studies have created genome-scale catalogues of DNase I hypersensitive sites (DHSs), which demark potentially functional regulatory DNA. To improve our understanding of the molecular evolution mechanism of antiviral differences between species, we leveraged 14,130 DHSs derived from 145 cell types to characterize the regulatory landscape of the TRIM region. Subsequently, we compared the alignments of the DHSs across six primates and found 375 DHSs that are conserved in non-human primates but exhibit significantly accelerated rates of evolution in the human lineage (haDHSs). Furthermore, we discovered 31 human-specific potential transcription factor motifs within haDHSs, including KROX and SP1, that both interact with HIV-1. Importantly, the corresponding haDHS was correlated with antiviral factor TRIM23. Thus, our results suggested that some viruses may contribute, through regulatory DNA differences, to organismal evolution by mediating TRIM gene expression to escape immune surveillance.

1. Introduction

Positive selection acting on antiviral restriction factors is consistent with virus–host interactions and reflects millions of years of an evolutionary arms race between the host and viral pathogens in primates [1,2]. The adaptive evolution in the tripartite motif-containing (TRIM) protein family is reminiscent of this prey–predator-like ‘arms race’ between virus and host. It is well known that the TRIM superfamily is emerging as a central component of the innate antiviral immunity of primates [3]. Recent studies have shown that increases in the ability of the TRIM family members to recognize or interact with the specific virus are beneficial to the host genome [4,5]. A previous study reported that gene expression differences rather than protein-coding changes may underlie some of the anti-retroviral immune mechanism [6]. For instance, some viruses employ strategies to downregulate TRIM gene expression to escape immune surveillance. The best example of this is the level of TRIM15 expression during HIV infection [6]. However, it is unclear the role of the non-coding region of TRIM family members in preventing viral replication and its contribution to the differences among primates in viral infection.

The human genome contains as many as 100 genes in the TRIM multigene family [7]. Although the TRIM multigene family is recognized for encoding several bona fide restriction factors [7], still most TRIM family members remain functionally uncharacterized, along with their potential for encoding antiviral activities [8]. Many recent works have shown that under long-term or recurrent viral selection pressure, a substantial number of the TRIM family members, such as TRIM5, TRIM15, TRIM21, TRIM22, TRIM25, TRIM51 and TRIM38, mediating antiviral
activities by directly targeting viral proteins, have been shown to have characteristic signatures of positive selection act on the gene encoding regions. For example, TRIM5a, the most well-studied TRIM members that provide a post-entry defence against retroviral infection [5,9], has been reported its signatures of positive selection were discovered in coiled-coil and B30.2 protein domain [9]. However, some TRIM members such as TRIM19/PML confers resistance to a broad range of DNA and RNA viruses [9], it did not show a signature of positive selection acting on the gene’s coding region [7]. Despite these interesting findings in the TRIM family coding region, little is known about the specific regulatory sequence variation of the TRIM multi-gene family that is responsible for changes in transcription and phenotype, thus we made a bold speculation that regulatory sequence variation may play an important role in the evolution of TRIM members. Therefore, the aim of this study was to improve our understanding of the molecular evolution mechanism of antiviral differences between species.

Tremendous technological progress over the last decade has resulted in the rising number of completely sequenced primate genomes and has led to a vast catalogue of the variation in the genomes and epigenomes across many primates [10,11]. However, interpreting the evolutionary, functional and host–pathogen interactions significance of these differences and identifying what regions of the genome are responsible for human-specific phenotypes remains a formidable challenge [11]. Furthermore, in our previous study, our results showed that variation in non-coding regulatory elements is a considerably important driver of gene expression differences and phenotypic differences, as originally proposed four decades ago [11]. To the best of our knowledge, the DNase I assay has proven to be an elegant and powerful methodology for the discovery of in vivo regulatory sequences in complex genomes [12,13]. DNase I preferentially cleaves the regions of open and active DNA, making it a highly successful and extensively validated assay to identify regulatory elements, including promoters, enhancers, repressors, boundary elements and locus control regions [10,11]. As part of the ENCODE Project and Roadmap Epigenomics Project, extensive maps of DNase I hypersensitive sites (DHSs) have been created in over 140 cell types and high-resolution DNase I footprints in over 30 cell types [14].

In this study, we analysed DHS in the TRIM region identified in the Encode Project [12] and fibroblasts of human, chimpanzee and macaque beside lymphoblastoid cell lines (LCLs) of human and chimpanzee from Crawford et al. [10]. As a result, we identified thousands of DHSs that are conserved in non-human primates and hundreds of DHSs that exhibit significantly accelerated rates of evolution in the human lineage only (haDHSs). We performed a series of bioinformatics analyses to better understand the functional and biological characteristics of haDHSs in the TRIM regulatory region, especially the association with virus. These studies have greatly improved our understanding of aspects of the non-encoding regulatory sequence, transcription factor and target genes in primates.

2. Results and discussion
(a) Identification of regulatory DNA in the TRIM family region
To better define regulatory DNA in the TRIM family region that has been subject to human-specific adaptive evolution, we leveraged experimentally defined maps of DHSs from 130 cell types identified in ENCODE and 15 cells were from Crawford et al. [10]. First, we obtained the coordinates of 76 human (Homo sapiens) TRIM genes from the NCBI database. Subsequently, we converted the coordinates of the TRIM superfamily to hg19 with liftOver, using the University of California Santa Cruz (UCSC) Genome Browser website (http://genome.ucsc.edu/). After merging DHSs across different primate species and finding non-coding regulatory DNA in the TRIM family region, we used the Bedops package [15]. In total, we detected 193,409 DHSs in the TRIM superfamily region across 145 cell types. Among of these DHSs, 140,154 were discovered from Crawford et al. [10]. Subsequently, we compared DHSs across three species. As a result, we identified 26,474, 31,724 and 27,138 DHSs in the TRIM family region from human, chimpanzee and macaque fibroblasts, respectively. Furthermore, we discovered 16,561 human-specific DHSs, 20,880 chimpanzee-specific DHSs, 17,069 macaque-specific DHSs and 11,239 common DHSs (figure 1) by using stringent criteria (50% quantile) for overlap across three species. We speculated that these species-specific DHS sites may be responsible for the varied biological functions across species during primate evolution.

(b) Positive selection within the TRIM multigene family in primates
It is well known that TRIM proteins are a large family of proteins that have been implicated in many biological processes, including cell differentiation, apoptosis, transcriptional regulation and signalling pathways [16], particularly because of their roles in virus infection [17]. In this study, based on the methods used in the previous studies of DHS test for positive selection [11], we did three likelihood ratios tests (see Methods) on those DHSs that were located in the TRIM family regulatory region in primate genomes. The first ‘CON&ACC’ likelihood ratio test (LRT) tested for distinguishing the evolutionary rate across all six primates in the DHS compared with the neutral sequence. The second ‘ACC’ LRT tested for distinguishing the evolutionary rate of DHS only in human lineage compared with the neutral sequence. The last ‘CON’ LRT tested for the DHS that converged in other five primates except human. The result of these tests were evaluated by p-value, if the absolute value of the p-value is less than 0.05, then we considered this DHS is significantly accelerated or conversed. Notably, in the ‘CON&ACC’ test, the p-value ranged from −1 to 1, as the negative p-value indicates accelerated. As a result, we found strong signatures of episodic positive selection activity on genes non-encoding regulatory regions in the primate TRIM family members. More specifically, of these DHSs, 225 were accelerated across all six primates, 3774 were conserved across all six primates, 4412 were conserved across all five primates except human (figure 2 and electronic supplementary material, figure S1), and 833 evolved rapidly on the human lineage (figure 2 and electronic supplementary material, figure S1). Notably, we identified 375 haDHSs based on the ‘ACC’ and ‘CON’ test. The result indicates that most regulatory sequences have been selectively maintained through millions of years of primate evolution, suggesting a role in important regulatory function. By contrast, only few DHSs are rapidly evolved in human lineage and conserved in other non-human primates, suggesting these DHSs may contribute to the different regulatory function of the TRIM family non-coding region between primates.
Additionally, to investigate the diversity of haDHSs in the human genome, we searched all single nucleotide polymorphisms (SNPs) in the haDHSs identified by us from the 1000 genomes project (http://grch37.ensembl.org/Homo_sapiens/Info/Index). Our result indicated most of SNPs within haDHSs had low allele frequency, but still there was difference among different populations, especially the allele frequency of these SNPs had a great difference between European and African populations (figure 3), suggesting haDHSs were under selection pressure.

Next, we investigate the biological function of these haDHSs and their correlation with the TRIM family. According to the DHS-target gene data [11,18], we discovered that 22 TRIM target genes (electronic supplementary material, table S1), including TRIM2, TRIML2, TRIM3, TRIM6, TRIM8, TRIM11, TRIM13, TRIM14, TRIM15, TRIM17, TRIM23, TRIM26, TRIM33, TRIM34, TRIM36, TRIM39, TRIM40, TRIM44, TRIM45, TRIM46, TRIM55 and TRIM68, where the
corresponding DHSs were accelerated in humans. In 18 TRIM target genes, including TRIM9, TRIM10, TRIM16, TRIM21, TRIM24, TRIM58, TRIM42, TRIM50, TRIM54, TRIM59, TRIM60, TRIM63, TRIM65, TRIM66, TRIM69, TRIM71, TRIM74 and TRIM77, the corresponding DHSs were conserved across all five primates except humans, including chimpanzee, gorilla, orangutan, macaque and marmoset (electronic supplementary material, table S1). Owing to the limitations of current research, the majority of TRIM genes, along with their DHSs, remain largely uncharacterized. In this case, our TRIM DHS data cannot represent the actual whole DHSs of TRIM regulatory regions.

Surprisingly, the regulatory region of four antiviral host restriction factors, including TRIM8, TRIM11, TRIM13 and TRIM14, is undergoing positive selection, but there is no positive selection evidence in the protein coding region [10]. It has been recently reported that downregulation of TRIM11 in HEK299 or HeLa cells enhanced HIV-1 release, indicating that the endogenous protein contributed to the suppression [8]. Similarly, animal studies have shown the overexpression of the antiviral factor TRIM14 gene in human HIV-associated and monkey SIV-associated lymphomas [19,20]. Therefore, we suggested viruses may employ strategies to down modulate the expression of TRIM genes by means of the regulatory region to escape immune surveillance under long-term or recurrent viral selection pressure between the antiviral factors and the virus.

(d) Analysis of potential transcription factor binding sites in human lineage DNase I hypersensitive sites

To improve our understanding of how transcription factor binding differs between humans and five non-human primates, we next investigated transcription factor motifs in the sets of species-specific and rapidly evolving haDHSs. Data from five species motifs were used to classify regions as human gains or human losses. More specifically, we used the JASPAR and TRANSFAC database to define human binding site gains as cases where the sequence was a perfect match to the motif consensus sequence in the human branch only. Similarly, human-specific losses were defined as matching the consensus sequence in five non-human primates’ branches (chimpanzee, gorilla, orangutan, macaque and marmoset), all but the human branch. As a result, we detected 31 potential transcription factor binding sites (TFBSs) that were human-specific gained within 25 haDHSs, including potential transcription factor binding sites of 28 transcription factors (some transcription binding sites had the same corresponding transcription factors): PR, TRB2, Oct1, E2F1, AP2, TGIF, MZF1, CREB, HOXC13, NKX22, RPS8, SIX4, NEAT, KROX, SP1, MAZ, E47, HOXB9, Ets1, ELFI1, FLI1, SOX9, IRF, Oct4, PAX6, AP2GAMMA, PAX7 and FOXJ2 (table 1). However, none of the human loss motifs were found in the human-specific branch. Figure 4 shows one example of an E2F1 human-gained motif. We next made a comparison to find out whether these TFBSs were specifically found in the TRIM haDHS region or widely occurred in the rapidly evolved regulatory region. According to our study on DHSs in the major histocompatibility complex (MHC) regulatory region (not published yet), there were also a detection of the PR, SP1, IRF, AP1, AP2, FOXJ2, SREBP1, KROX, MAZR, HNF4, GATA2, AKEB6, OCT1, ELFI1, FLI1, MZF1, FOXD3, FOXO3A, RPS8, HNF3B, HOXC13 and AP4 binding sites within haDHSs in this region. However, we found that these TFBSs in the MHC rapidly evolved region did not act as human-specific gain binding sites (electronic supplementary material, table S2), providing strong evidence for the speculation that the human-specific TFBSs we identified were specific elements during the evolution of the TRIM regulatory region. However, SOX9 and GATA1 binding sites were detected both in TRIM and MHC haDHSs, and both act as a human-specific gain binding sites, indicating that SOX9 and GATA1 binding sites may undergo the selection pressure on a genome scale.

A significant fraction of regulatory variation can be attributed to changes in cis-regulatory sequences [21–23]. In order to find out what biological process these species-specific TFBSs may take part in, next we investigated the functions of the 28 transcription factors described above consistent with the identified transcription factors motifs, and it turned out they played important roles in cancer [24,25], viral response [26,27], cell proliferation and tissue-specific differentiation [28,29]. Strikingly, we identified almost 61% of transcription factor motifs that were involved in the response to viral infection, including TBR2 (T-box brain protein 2) [30], OCT1 [31], E2F1 [32], AP2 [33], TGIF [34,35], MZF1 [36], CREB [37–39], NEAT [40], KROX [41], SP1 [31,37], E47 [42], Ets1 [43], ELFI1 [44,45], FLI1 [46], IRF [47], Oct4 [48,49] and FOXJ2 [50,51] (electronic supplementary material, table S3). For example, TBR2 has been shown to play a pivotal role in HIV-associated immune activation through regulation of the activity of CD8+ T lymphocytes [26]. Oct-1 has been shown to repress HIV-1LTR promoter activity and its transactivation by Tat [51]. In addition, AP2 [52], SP1 [31], TGIF [34] and KROX [41] also have interaction with HIV viral proteins. In the electronic supplementary material, table S3, the detailed antiviral function of these 17 transcription factors had been shown. Moreover, both KROX and SP1 genes interact with HIV-1. Importantly, the corresponding haDHS was correlated with the antiviral factor TRIM23 (table 1). These data, combined with the target genes regulated by haDHSs, will provide us with further insights into the functional significance and potential regulatory roles of haDHSs.

Furthermore, we noted the most of SNPs that are located in the identified human-specific gained TFBSs had low allele frequencies, indicating that these human-gained TFBSs were fixed in the modern human-being (electronic supplementary material, table S4). Interestingly, there were four SNPs (rs2176073, rs1528693, rs16744110, rs16934057) of which the mutated allele was the same as the ancestor allele as well as the corresponding base of five non-human primates. This result provided strong evidence that these four TF binding sites were diverged from ancestor and non-human primates, then fixed in the modern human population, except the mutated allele frequency of SNP rs1528693 was oddly high. As noted above, it can be surmised that the human-specific transcription factor motifs that we found in these haDHSs may be responsible for the varied viral infection differences among primate species by changing the expression of TRIM family members. These analyses revealed that human-specific sequence differences in transcription factor motifs found within these haDHSs are linked with these transcription factors, suggesting that fine-tuning regulatory networks by tinkering with the sequences that govern the expression of regulatory proteins has been an important target of positive selection during human evolution. In short, our data provide substantial new insights into sequences that have experienced human-specific adaptive regulatory evolution.
<table>
<thead>
<tr>
<th>chromosome</th>
<th>start</th>
<th>end</th>
<th>target gene</th>
<th>TF motif(human gains or loss)</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1</td>
<td>118123038</td>
<td>118125757</td>
<td>TTF2</td>
<td>PR(human gain)</td>
</tr>
<tr>
<td>chr1</td>
<td>154785009</td>
<td>154787937</td>
<td>CLK2,CHRNB2</td>
<td>TBR2(human gain)</td>
</tr>
<tr>
<td>chr1</td>
<td>154809820</td>
<td>154813654</td>
<td>UBE2Q1,IL6R,SHEDT,TRD10,EFNA1,DCST2,RP11-30X12.11,TRIM46,</td>
<td>Oct1(human gain)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>FAM18B,HCN3,PKLR,RUSC1</td>
<td></td>
</tr>
<tr>
<td>chr1</td>
<td>231376077</td>
<td>231379035</td>
<td>C1orf174,TSNAX,CP11-17H4.2,CAZ929,DISC1</td>
<td>E2F1(human gain)</td>
</tr>
<tr>
<td>chr1</td>
<td>33763498</td>
<td>33765921</td>
<td>HPCAS100PBP,VARS,RP11-11703.2,RP4-803A2.2</td>
<td>AP2(human gain)</td>
</tr>
<tr>
<td>chr3</td>
<td>108907435</td>
<td>108908147</td>
<td>RP11-14807.2,RETNLB</td>
<td>TGF(human gain)</td>
</tr>
<tr>
<td>chr3</td>
<td>108106397</td>
<td>108108746</td>
<td>IFT57,BYH5,DZIP3,RP11-14807.2,RETNLB</td>
<td>MZF1(human gain)</td>
</tr>
<tr>
<td>chr3</td>
<td>140305567</td>
<td>140305915</td>
<td>no</td>
<td>CREB(human gain)</td>
</tr>
<tr>
<td>chr4</td>
<td>165634343</td>
<td>165634680</td>
<td>no</td>
<td>HOX13(human gain)</td>
</tr>
<tr>
<td>chr4</td>
<td>188814238</td>
<td>188814409</td>
<td>no</td>
<td>NKX22(human gain)</td>
</tr>
<tr>
<td>chr4</td>
<td>153870991</td>
<td>153872450</td>
<td>RP11-1140F12.2,RP11-285C1.2</td>
<td>RPS8(human gain)</td>
</tr>
<tr>
<td>chr4</td>
<td>188829042</td>
<td>188830050</td>
<td>RP11-1731C19.2</td>
<td>SIX4(human gain)</td>
</tr>
<tr>
<td>chr5</td>
<td>64481352</td>
<td>64483010</td>
<td>ADAMTS6</td>
<td>NFAT(human gain)</td>
</tr>
<tr>
<td>chr5</td>
<td>64919033</td>
<td>64921413</td>
<td>CENPK,PPWD1,15orf44,TRIM23,ERBB2IP</td>
<td>KROX,SP1,MAZ2(human gain)</td>
</tr>
<tr>
<td>chr6</td>
<td>30568140</td>
<td>30569070</td>
<td>HCG22,MRM</td>
<td>SP1(human gain)</td>
</tr>
<tr>
<td>chr7</td>
<td>103569120</td>
<td>103569270</td>
<td>RELN</td>
<td>E47(human gain)</td>
</tr>
<tr>
<td>chr7</td>
<td>138553840</td>
<td>13855690</td>
<td>no</td>
<td>HOXB9(human gain)</td>
</tr>
<tr>
<td>chr7</td>
<td>99996861</td>
<td>99997455</td>
<td>no</td>
<td>Ets1,ELF1,FLI1(human gain)</td>
</tr>
<tr>
<td>chr8</td>
<td>27062086</td>
<td>27063006</td>
<td>no</td>
<td>SOX9(human gain)</td>
</tr>
<tr>
<td>chr9</td>
<td>11965414</td>
<td>11965857</td>
<td>no</td>
<td>RPS8(human gain)</td>
</tr>
<tr>
<td>chr11</td>
<td>4711074</td>
<td>4712357</td>
<td>OR52B3P,OR51R1P,C11orf40,TRIM68,MMP26,OR51N1P,</td>
<td>SOX9(human gain)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>OR513P,OR51A4,OR51A2,OR52S5P,OR52A5</td>
<td></td>
</tr>
<tr>
<td>chr11</td>
<td>5312450</td>
<td>5312899</td>
<td>AC10438.9,OR51AB1P</td>
<td>IRF(human gain)</td>
</tr>
<tr>
<td>chr13</td>
<td>50652880</td>
<td>50654010</td>
<td>no</td>
<td>Otx4(human gain)</td>
</tr>
<tr>
<td>chr15</td>
<td>45399950</td>
<td>45419961</td>
<td>DUOX2A2,SLC30A4,B2M,SORC1,</td>
<td>PAX6,AP2GAMMA,PAX7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(human gain)</td>
</tr>
<tr>
<td>chr17</td>
<td>54516680</td>
<td>54519706</td>
<td>no</td>
<td>FOX02(human gain)</td>
</tr>
</tbody>
</table>
Analysis of transposable elements in different DNase I hypersensitive sites regions

Many regulatory regions were derived from transposable elements, such as DHS [53]. Therefore, we intend to find out whether the distribution of transposable elements associated with TRIM DHSs is different between haDHSs and con-DHSs (DHSs that conserved both in human and other five primates). We identified the transposable elements that overlapped with TRIM DHSs and our result showed the distribution of the SINE and LINE family between haDHSs and con-DHSs exhibited a significant difference (significance was examined with a chi-squared test, the \( p \)-values were 0.057 and 0.024, respectively; figure 5 and electronic supplementary material, figures S2 and S3). As previous studies have shown, the member of SINE family, MIR element putatively participated in regulating the host gene transcription [54] and showed the characteristics that were under selection during evolution [55]. Furthermore, the transcription of the SINE family member, Alu element increased when under stress conditions, such as virus infection [56]. Meanwhile, LINE-1 retrotransposition can lead to types of genomic instability [57] and a previous study assumed that the host processed to act to restrict the LINE-1 retrotransposition [58]. Taken together, the different distribution of SINE and LINE transposable elements provided evidence of the speculation that the transposable elements in TRIM haDHSs may play a much more important role when facing selective pressure such as viral infection during human evolution and contribute to the difference of antiviral ability between species through transcription regulation. In brief, our result showed the variant distribution of transposable elements associated with TRIM haDHSs and con-DHS, emphasizing the importance of transposable elements in DHSs during human evolution.

3. Methods

(a) DNase I hypersensitivity sites
We used DNase I hypersensitivity peaks previously published as part of the ENCODE [12,59] and Roadmap Epigenomics [60] projects. More information about the 130 cell types is available in the electronic supplementary material, table S5. All peaks were called, using the hotspot algorithm [61] and represent the 150 bp region of maximal DNase I signal. DHS data from another 15 cell types were obtained from Crawford et al. [10]. The cells were primary skin fibroblasts from three human, three chimpanzee and three macaque individuals, and B cells immortalized with the Epstein–Barr virus (EBV), also known as LCLs, were obtained from the same three human and three chimpanzee individuals. EBV does not reliably transfect macaque lymphocyte cells, so matched macaque LCLs were not available for this study. All of these data were converted to hg19 using the UCSC liftOver...
tool. For aggregate analyses over DHS across cell types, peak or footprint locations were merged across cell types, using the BEDOPS package [15].

(b) Primate alignments
We downloaded the six primate Enredo-Pecan-Ortheus Pipeline (EPO) alignments from Ensembl version 70 [62] and obtained an alignment for each DHS and the surrounding 50 kb of sequence. We masked all sites that were polymorphic in the 1000 Genomes project [63] integrated phase 1 data (March 2012) at less than 95% allele frequency, all repeat masked bases (lower case mark-up in the EPO alignment), and all sites that were part of a cytosine-phosphate-guanosine in any species in the alignment. In the surrounding 50 kb, we additionally masked all segmental duplications (UCSC table browser), coding exons (UCSC RefSeq genes) padded by 10 bp to remove splice sites, promoters (500 bp upstream of transcription start sites), other DHSs, and phastCons Eutherian mammal and primate conserved elements (UCSC phyloP46way). This helped ensure that the 50 kb surrounding region was a more appropriate approximation of the neutral evolutionary model for each DHS. We filtered any DHS in which either less than 90% of the bases remained unmasked in the DHS or less than 15 kb remained unmasked in any of the six primates in the neutral region. Note, the EPO alignment is based on GRCh37 (hg19), and all subsequent analyses were performed using GRCh37 coordinates. Given that we focused on conserved elements, which are by definition located in regions of the genome that are well resolved and alignable, we do not anticipate that realignment to GRCh38 would significantly affect our results.

(c) Identifying conserved and accelerated DNase I hypersensitive sites
DHSs that passed filtering were tested for overall conservation along the primate lineage with software from the PHAST package [64,65]. For each DHS, we first ran PHYLOFIT on the neutral alignment of the surrounding 50 kb with the parameters—rates 4—subst-mod SSREV—EM. We used the Newick tree provided with the six primate alignment in Ensembl. The resulting file was used as the neutral model while running PHYLOP; PHYLOP was run with the parameters—method LRT—mode CON after removing the human sequence from the alignment. DHSs that were conserved at a false discovery rate (FDR) of 1% as determined with the Q-value package (http://github.com/jdstorey/qvalue) for R (https://www.r-project.org/) were then tested for human acceleration. For this test, we used the same neutral model of evolution, this time using the parameters—method LRT—mode ACC—subtree homo_sapiens. DHSs significant for human acceleration at an FDR of 5% were considered in further analyses.

To determine the overall rate of evolution in the neutral regions compared with the haDHSs, we first concatenated sequences from both sets of regions and then conducted the same set of tests on the regions as a whole. To determine how much faster the human branch in the haDHSs was evolving compared with the expected rate, we multiplied the estimated neutral human branch length by the estimated conservation scale factor and divided the actual haDHS human branch length by this expected number.

(d) The population diversity of human lineage DNase I hypersensitive sites
We downloaded the SNPs data from 1000 genomes project. Basing the human reference genome, we identified the SNPs that located in our identified haDHSs, with the allele frequencies among populations and the polymorphic bases of SNPs. To investigate the diversity of these SNPs in the human genome, we performed a pairwise comparison with the SNP allele frequencies of different populations.

(e) Motif analysis
Protein binding motif locations were generated genome-wide with FIMO motif scanning software [66], version 4.6.1, using a p-value threshold of less than or equal to $1 \times 10^{-7}$. Motif models were obtained from TRANSFAC [67], v. 2011.1.

(f) Identification of transposable elements associated with DNase I hypersensitive sites
We downloaded the repeats annotation files from UCSC, and identified the transposable elements that overlapped with TRIM DHSs, defining as DHSs-associated transposable elements. Then, we counted the transposable elements in different DHS regions and made a comparison, using a chi-squared test to examine whether the difference of transposable elements between different DHS regions is significant.


Competing interests. We declare we have no competing interests.

Funding. This work was supported by joint funding from the School of Bioscience and Bioengineering, South China University of Technology and BGI-ShenZhen.

Acknowledgements. We thank the Crawford research group from Duke University. This work was supported by joint funding from the Education Department of Guangdong Province (2014KZDXM009).

Reference


