Scalable method for groupwise comparisons and its application to chromatin accessibility among primates

Authors and affiliations

Lee E. Edsall[1,2,3], Alejandro Berrio[4], William H. Majoros[5], Devjanee Swain-Lenz[4], Shauna Morrow[4], Yoichiro Shibata[1,2], Alexias Safi[1,2], Gregory A. Wray[1,4,*], Gregory E. Crawford[1,2,*], Andrew S. Allen[5,*]

[1]Center for Genomic and Computational Biology, Duke University, Durham, NC

[2]Department of Pediatrics, Division of Medical Genetics, Duke University, Durham, NC

[3]University Program in Genetics and Genomics, Duke University, Durham, NC

[4]Department of Biology, Duke University, Durham, NC

[5]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC

*Corresponding authors: E-mails: gwray@duke.edu; greg.crawford@duke.edu; andrew.s.allen@duke.edu

Data deposition: The project has been deposited at NCBI GEO under the accession number GSE129034

Abstract

Changes in transcriptional regulation are thought to be a major contributor to the evolution of phenotypic traits. We developed a new method to identify DNase I Hypersensitive (DHS) sites with differential chromatin accessibility between species. Our method replaces threshold-based multiple pairwise comparisons with a single quantitative test, which detects subtle differences while scaling well to multiple taxa. We applied this method to DHS sites in fibroblast cells from five primates (human, chimpanzee, gorilla, orangutan, and rhesus macaque). We identified

approximately 90,000 DHS sites, of which 59% are non differential between species, 27% are differential and likely due to a single evolutionary change, and 14% are differential due to multiple evolutionary changes. We found that using closely related species in our analyses allows us to distinguish between accessibility changes that are specific to a single species and those that have experienced multiple changes in chromatin accessibility during primate evolution. The non differential sites are enriched for nucleotide conservation, while the species-specific changes are enriched for positive selection. Differential DHS sites with decreased chromatin accessibility relative to rhesus macaque occur more commonly in proximal regulatory elements, while those with increased chromatin accessibility relative to rhesus macaque occur more commonly in distal regulatory elements. Differential DHS sites overlapping proximal regulatory elements display less cell-type specificity than those overlapping distal regulatory elements. Taken together, these results identify several classes of chromatin accessibility, each with distinct characteristics of selection, genomic location, and cell-type specificity.

Key words: cis-regulatory evolution, comparative functional genomics, positive selection, chromatin accessibility, transcriptional regulation

Introduction

It has long been hypothesized that phenotypic differences between species are more often due to genetic variation in non-coding regulatory regions than in protein-coding regions (King and Wilson 1975; Wray 2007; Wittkopp and Kalay 2011). The development of diverse genome-wide assays, combined with the publication of primate reference genomes, has allowed identification of inter-species differences in gene expression (Cáceres et al. 2003; Gilad et al. 2006; Blekhman et al. 2008; Brawand et al. 2011), DNA methylation (Pai et al. 2011; Zeng et al. 2012; Hernando-Herraez et al. 2013), histone modifications (Zhou et al. 2014; Villar et al. 2015),

47 transcription factor binding motifs (Dermitzakis and Clark 2002; Odom et al. 2007; Schmidt et
48 al. 2010), chromatin accessibility (Shibata et al. 2012; Gallego Romero et al. 2018), and
49 alternative splicing (Blekhman et al. 2010; Barbosa-Morais et al. 2012). These differences in
50 molecular function among primate species can provide valuable insights into species-specific
51 trait differences, including disease risk (Prabhakar et al. 2008; Boyd et al. 2015; Prescott et al.
52 2015).

53 Previous approaches to analyzing multi-species data have used multiple pairwise
54 comparisons to detect differences between species (Robinson et al. 2010; Love et al. 2014).
55 Applying these methods to comparisons between three or more groups becomes cumbersome
56 because it requires multiple pairwise comparisons followed by an overlap step that intersects the
57 results from those comparisons. The number of required pairwise comparisons increases
58 substantially as the number of groups is increased (e.g., 10 comparisons for 5 species).
59 Calculating p-values for multiple species comparisons is difficult because species are not
60 completely independent and indeed differ in their degree of dependence due to phylogenetic
61 relationships.

62 As an alternative to performing multiple pairwise comparisons, we developed a negative
63 binomial generalized linear model that jointly models data across multiple species and replicates,
64 and can be applied to any count-based data such as DNase-seq, RNA-seq, ChIP-seq, and ATAC-
65 seq. We applied this method to chromatin accessibility DNase-seq data from cultured skin
66 fibroblasts obtained from human, three recently diverged great apes (chimpanzee, gorilla, and
67 orangutan), and rhesus macaque for use as an outgroup. We identified a total of 89,744 DNase I
68 Hypersensitive (DHS) sites that are directly comparable between species, of which 36,666 (41%)
69 displayed a statistically significant difference between species. Using effect estimates from the

70 negative binomial generalized linear model, we inferred the direction of the change relative to
71 rhesus macaque and the possible branch or branches on which the changes occured.

72 **Materials and Methods**

73 ***DNase-seq Experiments and Sequencing***

74 Fibroblast cell lines from 15 individuals comprising 3 biological replicates from each of
75 five primate species (human, chimpanzee, gorilla, orangutan, and rhesus macaque) were obtained
76 from Coriell (**Supplementary Table 1**). It is estimated that human and chimpanzee diverged 7
77 million years ago, gorilla diverged from the human-chimpanzee ancestor 10 million years ago,
78 orangutan diverged from the human-chimpanzee-gorilla ancestor 18 million years ago, and
79 rhesus macaque diverged from the human-chimpanzee-gorilla-orangutan ancestor 30 million
80 years ago (Schrago and Voloch 2013). DNase-seq experiments were performed as previously
81 described (Shibata et al. 2012). DNase-seq libraries were generated from 50 million cells and
82 sequenced on Illumina instruments (**Supplementary Table 1)**.

83 ***DNase-seq Read Mapping and Conversion to Human Genome***

84 Due to the use of MmeI to generate DNase-seq libraries (Boyle et al. 2008), genomic
85 DNA fragments are only 20 bases long. Therefore, sequencing reads were trimmed to 20 bases
86 using a custom perl script. Reads for each species were mapped to its native genome: hg19 for
87 human, panTro4 for chimpanzee, gorGor3 for gorilla, ponAbe2 for orangutan, and rheMac3 for
88 rhesus macaque (Lander et al. 2001; Chimpanzee Sequencing and Analysis Consortium 2005;
89 Locke et al. 2011; Yan et al. 2011; Scally et al. 2012). Reads were mapped using Bowtie version
90 0.12.9 (Langmead et al. 2009) (parameters: -trim5 0 --trim3 0 -m 1 -l 20) as part of a custom
91 two-step pipeline. In the first step ("tier 1"), reads were required to match to a unique location

92 with no mismatches (parameter: -n 0). In the second step ("tier 2"), unmapped reads from step
93 one were re-mapped with a relaxed mismatch parameter of one mismatch (parameter: -n 1).
94 Reads that mapped to multiple locations or had more than one mismatch were discarded.
95 Samtools version 0.1.19-44428cd (Li et al. 2009) was used to convert the sam files from each
96 step to bam files, merge them into one file, and remove duplicate reads (defined as having the
97 same chromosomal coordinates). Bedtools v2.17.0 (Quinlan and Hall 2010) was used to convert
98 the bam files to bed files. Details on the number of reads in the input files and at each step are
99 included in **Supplementary Table 2**.

100 Reads from the non-human samples were converted from their native genomic
101 coordinates to hg19 coordinates using a three-step process that removed reads without a one-to-
102 one relationship between the genomes. In each step, read coordinates were converted from one
103 genome to the other using the UCSC liftOver software (Hinrichs et al. 2006) with a minMatch
104 parameter of 0.8, which requires that 80% of the read maps to the new genome. Note that this
105 parameter filters only on genomic coverage, not sequence identity. In the first step, read
106 coordinates were converted from their native genome to hg19. Read coordinates that successfully
107 lifted to hg19 were then lifted back to the native genome. Read coordinates that did not lift back
108 to the same coordinates on the native genome were removed. Reads that did lift back to the same
109 coordinates were lifted back to hg19 for further processing. An additional filtering step was
110 added to ensure the reads did not map to a duplicated region. In that step, overlapping reads on
111 the native genome were merged into a region, and that region was lifted to hg19. Regions that
112 failed to lift uniquely to hg19 were flagged and reads that overlapped them were removed.
113 Because some of the samples were from males and some were from females (**Supplementary**
114 **Table 1**), we removed reads that mapped or lifted over to the human X or Y chromosomes to

115 eliminate any sex-specific bias. Details on the number of reads lifted over and remaining after
116 removal of sex chromosomes are included in **Supplementary Table 2**. Phylogenetic trees were
117 drawn with ggtree (Yu et al., 2016).

118 ***DHS Site Identification and Filters***

119 To avoid bias due to large differences in depth of library sequencing, 20 million reads
120 were randomly selected from samples with library sizes greater than 20 million reads. To
121 determine the appropriate subset size, we generated subsets of 10, 15, 20, and 25 million reads
122 and compared the average number of per-species DHS sites in each subset (**Supplementary**
123 **Figure 1**).

124 Peak calling was performed on each sample separately using the MACS2 callpeak
125 command with an FDR cutoff of 5% (Zhang et al. 2008) (version 2.1.0.20150420; parameters: --
126 nomodel --extsize 20 --qvalue .05). For each species, DHS sites were identified by taking the
127 union set of peaks that were found in at least 2 out of 3 biological replicates. We removed DHS
128 sites that overlapped the ENCODE blacklist
129 (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wgEncode
130 DacMapabilityConsensusExcludable.bed.gz) (Rosenbloom et al. 2013). To generate a master set
131 of DHS sites to use for cross-species comparisons, we first took the union set of DHS sites
132 identified in each species, then applied two filters. In the first filter, we removed DHS sites
133 without at least 95% genomic coverage between the start and stop coordinates for each of the
134 species. Genomic coverage was determined using the Multiz Alignment MAF file
135 (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/multiz100way/maf/) from the UCSC Genome
136 Browser Multiz Alignment of 100 Vertebrates (Blanchette et al. 2004), the Galaxy MAF
137 Coverage Stats tool at usegalaxy.org (Afgan et al. 2018), and custom perl scripts. Next, we

138 assigned read counts to each DHS site using bedtools v2.17.0 (Quinlan and Hall 2010). In the
139 second filter, we removed DHS sites without DNase-seq sequence reads in at least two biological
140 replicates from each species because they may be indicative of regions that cannot be uniquely
141 aligned. In other words, we expect at least some level of background DNase cutting across the
142 genome. See **Supplementary Table 3** for DHS site counts before and after each filtering step.

143 ***Principal Components Analysis***

144 We performed a principal components analysis on the read counts from the 15 samples
145 **(Supplementary Figure 4)**. We first normalized the counts by library size then ran the R
146 prcomp function with the center and scale parameters set to true. We also performed a principal
147 components analysis on the read counts from the 15 samples and 3 additional chimpanzee
148 samples from Pizzollo et al. 2018 **(Supplementary Figure 4)**. See **Supplementary Table 1** for
149 details on the samples.

150 ***Differential Site Identification and Classification***

151 The read counts for each DHS site were used as input to a custom R script that identified and
152 classified differential DHS sites. To address the over-dispersion problem inherent in count-based
153 sequencing data, we used the R package DSS (Wu et al. 2013) to calculate a dispersion
154 parameter for each DHS site, as well as a normalization offset (based on total library size) for
155 each sample. For each DHS site, the read counts, dispersion parameter, and normalization offset
156 were fit using a negative binomial generalized linear model with a log link function. Specifically,
157 we fit two models: a species informed model and a null model in which species was not
158 predictive of normalized counts. The species informed model modelled the expected counts by
159 $log\ (\lambda_j) = \alpha_j + x_j{}^T\beta$, where λ represents expected counts and j indexes the sample. The design

160 vector $\boldsymbol{x_j}$ indicates to which species the jth sample belongs and $x_j{}^T$ denotes its transpose. This
161 vector has 4 elements comprised of indicator functions of whether the sample is human,
162 chimpanzee, gorilla, or orangutan. Because accessibility changes are relative to rhesus macaque,
163 it is used as the intercept in the models and does not have an indicator function. Specifically,
164 $x_j{}^T = (1\ 0\ 0\ 0)$ if the jth sample is human; $x_j{}^T = (0\ 1\ 0\ 0)$ if the jth sample is chimpanzee;
165 $x_j{}^T = (0\ 0\ 1\ 0)$ if the jth sample is gorilla; and $x_j{}^T = (0\ 0\ 0\ 1)$ if the jth sample is orangutan. The
166 vector $\beta = (\beta_h, \beta_c, \beta_g, \beta_o)^T$ parameterizes the change in expected counts between species and
167 $\alpha_j$is a normalization offset for sample j. The null model assumes the vector $\beta$is zero and models
168 the expected counts by $log\ (\lambda_j) = \alpha_j$. These models were fit using the R package glm using
169 negative.binomial as the family. The DSS normalization offset value was used for the offset
170 parameter. The inverse of the DSS dispersion parameter value was used for the theta parameter.
171 The difference in deviances between these two models was used to form a likelihood ratio test of
172 whether the sites were differential. A Benjamini-Hochberg correction was performed using the R
173 function p.adjust. DHS sites with a corrected p-value of less than .01 were classified as
174 differential.

175 To determine which species (or combination of species) were different from rhesus
176 macaque, 15 contrasts were constructed using the β values estimated in the regression model
177 detailed above. Five contrasts were used to identify changes in a single species, six for changes
178 in two species, and four for changes in three species (**Supplementary Table 4**). Note that the
179 contrast for changes in rhesus macaque will also identify changes that occurred in the human-
180 chimpanzee-orangutan-gorilla internal branch. A value for each contrast c using constraint
181 matrix C and variance-covariance matrix $Var(C\beta)$ was calculated as
182 $c = [C\beta]^T[Var(C\beta)]^{-1}C\beta$, where $[Var(C\beta)]^{-1}$ denotes the matrix inverse of $Var(C\beta)$. A p-

183 value was calculated for each contrast using a chi-squared test and the contrast with the lowest p-
184 value was chosen. For 30 DHS sites, none of the p-values were less than .01, so the change was
185 marked as "other". The sign of the β was used to determine whether each change was an increase
186 (referred to as "increased accessibility") or decrease (referred to as "decreased accessibility")
187 relative to rhesus macaque.

188 Note that since we classified only the sites that were shown to be differential and we
189 controlled the type one error rate in the differential site identification step, there is no need to
190 further adjust our analyses for the multiple contrasts being considered here. In the statistical
191 literature this is referred to as a "gateway" procedure as the differential site identification is a
192 gatekeeper for site classification.

193 We have included supplementary files that contain the input data for the R script
194 (**glm_analysis.input_file.txt**), the results from all of the analyses discussed in this paper
195 (**glm_output_and_analsyis_results.txt**), and the field information for the input and output files
196 (**input_and_output_file_field_information.xlsx**). Raw fastq files and bed files containing hg19
197 coordinates for the reads are available under GEO accession GSE129034. The reads, DHS sites,
198 and differential analysis classifications are available in a UCSC Genome Browser session at
199 http://genome.ucsc.edu/s/ledsall/2019primate. All scripts for the data processing and analyses
200 described above are available at http://github.com/ledsall/primate.

201 ***Testing for Positive Selection and Determining Vertebrate Conservation***

202 We performed selection analysis on both the differential and non differential DHS sites.
203 We tested for selection on the species branches for human, chimpanzee, gorilla, and orangutan,
204 and on the internal branches for human-chimpanzee and human-chimpanzee-gorilla.

205 As the stochasticity of the evolutionary process may be elevated in short alignments, we
206 expanded each DHS site that was smaller than 300 bases up to 300 bases, while maintaining the
207 size of any DHS site longer than 300 bases. We removed any sites that couldn't be expanded due
208 to gaps in the non-human genomes.

209 To investigate the extent of positive selection among the DHS sites, we used a branch-
210 specific method we first developed in 2007 (Haygood et al. 2007) and recently improved (Berrio
211 et al. under review). Briefly, the method uses a likelihood ratio test based on the maximum
212 likelihood estimates obtained from HyPhy (Pond et al. 2005). The branch of interest (e.g. human
213 species branch) is used as the foreground and the rest of the tree is used as the background. The
214 assumption for the background is the same for both the null and alternative models; specifically,
215 neutral evolution and negative (purifying) selection are permitted, but positive selection is not. In
216 the null model, the assumption for the foreground is the same as the one for the background. In
217 the alternative model, all three models of evolution are permitted (neutral evolution, negative
218 selection, and positive selection) in the foreground. This method is highly sensitive and specific
219 in that it can differentiate between positive selection and relaxation of constraint.

220 The method requires a 3kb reference alignment for each species that is used as a
221 putatively neutral proxy for computing substitution rates. To generate this alignment, we first
222 identified a set of functional regions on the human genome using annotations from the ENCODE
223 project at UCSC (http://genome.ucsc.edu/encode/downloads.html) (ENCODE Project
224 Consortium 2012) and annotations from the HoneyBadger2-intersect dataset from the ENCODE
225 and Roadmap Epigenomics projects
226 (https://personal.broadinstitute.org/meuleman/reg2map/HoneyBadger2-intersect_release/)
227 (Roadmap Epigenomics Consortium et al. 2015). We used the set of 56,893 putative promoter

228 regions, 1,598,323 putative enhancer regions, and 31,255 putative dyadic regions. We then
229 masked the genomes using those functional regions, along with 5' and 3' UTRs, coding and non-
230 coding RNAs, CpG repeats, microsatellite repeats, and simple repeats. Next, we extracted
231 windows of 300 bases and excluded those with substitution rates that are too high or slow
232 relative to the entire tree. Finally, we concatenated the set of these windows until we reached a
233 length of 3kb (Berrio et al. under review).

234 We used the PHAST library msa_split (Hubisz et al. 2011) to extract query regions from
235 the UCSC Genome Browser Multiz Alignment of 100 Vertebrates
236 (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/multiz100way/maf/) (Blanchette et al. 2004)
237 for the human, chimpanzee, gorilla, orangutan, and rhesus macaque genomes.
238 For each DHS site (called a query site), we used HyPhy (Pond et al. 2005) to fit the null and
239 alternative models and generate maximum likelihood values. We used a custom R script to
240 compute the likelihood ratio, which was used as a test statistic for a chi-squared test with one
241 degree of freedom to calculate a p-value. We classified a DHS site as under positive selection if
242 the p-value was less than .05. We were unable to successfully run HyPhy on 12 sites due to
243 unknown reasons and removed these regions from analysis.

244 We found that the substitution rates of the orangutan branch and the internal branches for
245 human-chimpanzee and human-chimpanzee-gorilla were extremely variable and caused an
246 overestimation of selection in these branches and in the subset of non functional data. Therefore,
247 we excluded these branches from the subsequent analyses and focused on the human,
248 chimpanzee, and gorilla branches.

249 We then tested for significant enrichments of positive selection in different classes of
250 DHS sites. The classes we investigated are 1) human-specific accessibility increases and

251 decreases; 2) chimpanzee-specific accessibility increases and decreases; 3) gorilla-specific
252 accessibility increases and decreases; and 4) non differential sites. We performed Fisher's exact
253 tests using a test statistic of the number of DHS sites classified as under positive selection. For
254 the human-specific accessibility changes, we performed two tests; 1) the human species branch
255 as the foreground compared to the chimpanzee species branch as the foreground; and 2) the
256 human species branch as the foreground compared to the gorilla species branch as the
257 foreground. For the chimpanzee-specific accessibility changes, we performed two tests; 1) the
258 chimpanzee species branch as the foreground compared to the human species branch as the
259 foreground; and 2) the chimpanzee species branch as the foreground compared to the gorilla
260 species branch as the foreground. For the gorilla-specific accessibility changes, we performed
261 two tests; 1) the gorilla species branch as the foreground compared to the human species branch
262 as the foreground; and 2) the gorilla species branch as the foreground compared to the
263 chimpanzee species branch as the foreground. For the non differential sites, we performed three
264 tests; 1) the human species branch as the foreground compared to the chimpanzee species branch
265 as the foreground; 2) the human species branch as the foreground compared to the gorilla species
266 branch as the foreground; and 3) the chimpanzee species branch as the foreground compared to
267 the gorilla species branch as the foreground. We used a Bonferroni correction to adjust for the
268 multiple tests performed.

269 To visualize the strength of selection, we computed the statistic ζ(zeta), representing the
270 ratio of evolution, by calculating the ratio of the substitution rate in each query compared to the
271 reference alignment; we computed ζ for the human, chimpanzee, and gorilla species branches.
272 This parameter is homologous to ω (omega), the ratio of dN/dS, where a value of $\omega < 1$

273 indicates constraint or negative selection; a value of $\omega = 1$ indicates neutrality; and a value of
274 $\omega > 1$ indicates positive selection.

275 We then tested whether the distributions of ζ differed between classes of differential DHS
276 sites. The classes we investigated are 1) human-specific increased accessibility; 2) chimpanzee-
277 specific increased accessibility; 3) gorilla-specific increased accessibility; 4) orangutan-specific
278 increased accessibility; 5) human-chimpanzee increased accessibility; 6) human-chimpanzee-
279 gorilla increased accessibility; 7) human-specific decreased accessibility; 8) chimpanzee-specific
280 decreased accessibility; 9) gorilla-specific decreased accessibility; 10) orangutan-specific
281 decreased accessibility; 11) human-chimpanzee decreased accessibility; and 12) human-
282 chimpanzee-gorilla decreased accessibility. We performed Wilcoxon tests on human-specific
283 increased accessibility against the other classes of DHS sites with increased accessibility (classes
284 2-6) and used a Bonferroni correction to adjust for the multiple tests. Similarly, we performed
285 Wilcoxon tests on human-specific decreased accessibility against the other classes of DHS sites
286 with decreased accessibility (classes 8-12) and used a Bonferroni correction to adjust for the
287 multiple tests. Finally, we performed a Wilcoxon test on the non differential sites against the non
288 functional sites defined above.

289 To determine the amount of vertebrate conservation, we computed the median value of
290 PhastCons scores for each DHS site using bedops (Neph et al. 2012), the UCSC 100-way
291 PhastCons table (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/phastCons100way) (Siepel et
292 al. 2005; Pollard et al. 2010), and custom scripts. The PhastCons score represents the probability
293 of a particular base being conserved. The values range from 0 to 1, with higher values
294 representing an increased probability of conservation. We classified a DHS site as constrained if
295 the median PhastCons score was above 0.9.

296 We then investigated whether the amount of conservation differed between differential
297 and non differential DHS sites. We used the percentage of constrained DHS sites as our test
298 statistic and performed a Fisher's exact test on non differential sites compared to three classes of
299 differential DHS sites; specifically, 1) human-specific accessibility changes; 2) chimpanzee-
300 specific accessibility changes; and 3) gorilla-specific accessibility changes. We used a
301 Bonferroni correction to adjust for the multiple tests.

302 ***Intersection with Human Putative Regulatory Annotations***

303 We characterized each DHS site as a proximal element, distal element, or unannotated
304 region using the HoneyBadger2-intersect dataset from the ENCODE and Roadmap Epigenomics
305 projects (https://personal.broadinstitute.org/meuleman/reg2map/HoneyBadger2-
306 intersect_release/) (Roadmap Epigenomics Consortium et al. 2015). We used the putative
307 promoter and enhancer regions as above, but did not use the putative dyadic regions. We used
308 bedtools2 (Quinlan and Hall 2010) to identify DHS sites that overlapped the annotated promoters
309 (which we characterized as "proximal elements") and enhancers (which we characterized as
310 "distal elements"). DHS sites that didn't overlap promoters or enhancers were characterized as
311 unannotated regions.

312 ***Determining Cell-type Specificity***

313 We characterized the cell-type specificity of each DHS site by intersecting it with DHS
314 sites from 125 human cell types and tissues (Thurman et al. 2012). We used bedtools2 (Quinlan
315 and Hall 2010) to find the overlap with the wgEncodeAwgDnaseMasterSites dataset from the
316 UCSC Genome Browser
317 (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeAwgDnaseMasterSites

318 /wgEncodeAwgDnaseMasterSites.bed.gz) . We assigned a score to each DHS site representing
319 its cell-type specificity. The score was calculated as $1 - \frac{N}{126}$, where *N* represents the number of
320 cell-types and tissues in which the DHS site is present (including the fibroblast cell line from this
321 study). The score ranges from 0 for a DHS site present in all tissues and cell types to $\frac{125}{126}$ (which
322 is approximately 0.99) for a DHS site present in only our dataset.

323 We then asked whether the distribution of cell-type specificity scores varied between
324 different classes of DHS sites. We subset the DHS sites into those overlapping proximal
325 elements and those overlapping distal elements. For each subset, we investigated the following
326 classes: 1) human-specific increased accessibility; 2) chimpanzee-specific increased
327 accessibility; 3) gorilla-specific increased accessibility; 4) orangutan-specific increased
328 accessibility; 5) human-chimpanzee increased accessibility; 6) human-chimpanzee-gorilla
329 increased accessibility; 7) human-specific decreased accessibility; 8) chimpanzee-specific
330 decreased accessibility; 9) gorilla-specific decreased accessibility; 10) orangutan-specific
331 decreased accessibility; 11) human-chimpanzee decreased accessibility; 12) human-chimpanzee-
332 gorilla decreased accessibility; and 13) non differential sites.

333 **Results**

334 ***Method Development to Identify and Classify Differential DNase I Hypersensitive Sites Across***
335 ***Multiple Primate Species***

336 We developed a negative binomial generalized linear model based method to allow us to
337 quantitatively compare DNase I hypersensitive (DHS) sites across five primate species with one
338 statistical test, rather than performing multiple pairwise comparisons. This greatly simplifies the

339 analysis, scales well for additional replicates and species, and has straightforward p-value
340 calculations. The output of this test indicates for each DHS site whether or not at least one
341 species displays a statistically significant difference in read counts. In addition, the output from
342 the model includes the vector $\beta = (\beta_h, \beta_c, \beta_g, \beta_o)^T$, representing the changes in expected counts
343 compared to rhesus macaque for human, chimpanzee, gorilla, and orangutan respectively. We
344 used both the value and the sign (positive vs. negative) of the β values to classify the type of
345 difference between a given species and the outgroup, rhesus macaque (see Materials and
346 Methods). DHS sites that are not differential have values near zero for all of the β values **(Figure**
347 **1A)**. DHS sites that have increased accessibility in a particular species have a positive β for that
348 species with the other species having β values near zero (**Figure 1B** for human-specific
349 increased accessibility; **Supplementary Figure 2** for other species). DHS sites that have
350 decreased accessibility in a particular species have a negative β value for that species with the
351 other species having β values near zero (**Figure 1C** for human-specific decreased accessibility;
352 **Supplementary Figure 2** for other species). We also detect regions that have altered β values in
353 more than one species (**Figure 1D**, **Supplementary Figures 2 and 3**).
354 Using this approach, we identified 89,744 total DHS sites that can be compared across all
355 5 species at 1:1:1:1:1 orthologous genomic regions (Materials and Methods). As a first step in
356 analyzing these data, we carried out a principal components analysis and found that the first
357 principal component separated the single old world monkey (Rhesus macaque) from the four
358 great apes, while the second principal component recapitulated the phylogeny of the great apes
359 (**Supplementary Figure 4**). Because we are drawing on the original data from Shibata et al.
360 2012 for three species and have generated new data for two species (**Supplementary Table 1**),
361 we also investigated whether batch effects would overwhelm the species signal by comparing

362 principal components analyses with and without three additional chimpanzee samples generated
363 more recently from Pizzollo et al. 2018 (**Supplementary Table 1**). As shown in **Supplementary**
364 **Figure 4**, the Pizzollo et al. chimpanzee samples cluster with the original Shibata et al.
365 chimpanzee samples across the first four principal components (cumulative proportion of
366 variance of 0.53), suggesting that biological signal is retained even when samples are prepared
367 and sequenced years apart.

368 Of the 89,744 total DHS sites, 53,078 (59%) are not statistically significantly different
369 between species, 23,926 (27%) display a difference that likely resulted from a single
370 evolutionary event, and 12,710 (14%) that display a difference due to multiple evolutionary
371 events **(Figure 1E, Table 1)**. Of the 23,926 DHS sites with changes that are likely due to a
372 single event, about half (11,879) appear to have changed more recently and are detected only in a
373 single species (human, chimpanzee, gorilla, or orangutan) **(Figure 1F, Table 1)**. About one
374 quarter of single event changes (5,738) appear to have occured in either the human-chimpanzee
375 internal branch or the human-chimpanzee-gorilla internal branch **(Figure 1F, Table 1)**. The
376 remaining single events (6,309) occurred either recently in rhesus macaque or longer ago in the
377 internal human-chimpanzee-gorilla-orangutan internal branch **(Figure 1F, Table 1)**. Because we
378 are using rhesus macaque as the outgroup, we are unable to differentiate between changes in the
379 rhesus macaque species branch and changes in the human-chimpanzee-gorilla-orangutan internal
380 branch (see Discussion). Consistent with our earlier study (Shibata et al. 2012), as well as studies
381 by other groups (Reilly et al. 2015; Villar et al. 2015; Emera et al. 2016), the majority of the
382 changes are increased accessibility rather than decreased accessibility (see Discussion). For
383 changes on the species branches, there are approximately 10x the number of DHS sites with
384 increased accessibility as DHS sites with decreased accessibility; 12:1 for human, 8:1 for

385 chimpanzee, 11:1 for gorilla, 6:1 for orangutan. For changes on the internal branches, the ratio of
386 DHS sites with increased accessibility to DHS sites with decreased accessibility is much less; 2:1
387 for the human-chimpanzee branch, and nearly equal for the human-chimpanzee-gorilla branch
388 **(Figure 1G, Table 1)**.

389 ***Changes in Chromatin Accessibility Detected in a Single Species***

390 Using the methods described above, we identified 10,574 DHS sites with increased
391 accessibility specific to a single species; 2,539 in human, 1,539 in chimpanzee, 2,737 in gorilla,
392 and 3,759 in orangutan **(Figure 2A, Table 1)**. We identified 1,305 DHS sites with decreased
393 accessibility specific to a single species; 209 in human, 196 in chimpanzee, 254 in gorilla, and
394 646 in orangutan **(Figure 2B, Table 1)**. Heatmap overviews **(Figure 2A, 2B)** of each class of
395 increased accessibility and decreased accessibility show that these differences are not binary, but
396 instead span the continuum from extremely large differences to those that represent more modest
397 changes. Representative screenshots of individual genomic loci are provided **(Figure 2A, 2B)**.
398 A complete list of coordinates for all DHS sites, β values, and p-values that allow regions to be
399 filtered at different cutoff stringencies are included in **Supplementary File**
400 **glm_output_and_analsyis_results.txt**.
401 Even though we can't classify rhesus macaque-specific changes, we can identify sites
402 where rhesus macaque is different from the other four species. We identified 4,992 sites that
403 have increased accessibility in rhesus macaque relative to human, chimpanzee, gorilla, and
404 orangutan **(Figure 2A, Table 1)**. We identified 1,317 sites that have decreased accessibility in
405 rhesus macaque compared to human, chimpanzee, gorilla, and orangutan **(Figure 2B, Table 1)**.

406 ***Changes in Chromatin Accessibility that Likely Occurred on Internal Branches***

407 Our method allows us to identify ancient changes in chromatin accessibility that likely
408 occurred as a single change on internal branches. Using rhesus macaque as an outgroup, we can
409 classify whether these internal changes are accessibility increases or decreases. We identified
410 1,735 DHS sites with increased accessibility and 814 DHS sites with decreased accessibility that
411 likely occurred during the common lineage of human and chimpanzee **(Figure 3, Table 1)**. We
412 identified 1,736 DHS sites with increased accessibility and 1,453 DHS sites with decreased
413 accessibility that likely occurred before the split between human, chimpanzee, and gorilla
414 **(Figure 3, Table 1)**.

415 ***Multiple Changes in Chromatin Accessibility***

416 In addition to detecting likely single changes in chromatin accessibility on either species
417 **(Figure 2)** or internal branches **(Figure 3)**, we also identified changes in chromatin accessibility
418 that appear to have occurred multiple times, resulting in different combinations of chromatin
419 accessibility patterns between species. There are many possible ways these differences could
420 have happened and our method cannot determine if these changes resulted from multiple
421 increases in accessibility, multiple decreases in accessibility, or a combination of increases and
422 decreases (see Discussion).

423 We identified 6,339 DHS sites where two species display increased accessibility relative
424 to rhesus macaque; 997 in human and gorilla; 997 in human and orangutan; 1,568 in chimpanzee
425 and gorilla; 657 in chimpanzee and orangutan; and 2,120 in gorilla and orangutan **(Figure 4A,**
426 **Table 1)**. We identified 1,801 DHS sites where two species display decreased accessibility
427 relative to rhesus macaque; 221 in human and gorilla; 424 in human and orangutan; 387 in

428 chimpanzee and gorilla; 252 in chimpanzee and orangutan; and 517 in gorilla and orangutan
429 **(Figure 4B, Table 1)**.
430 There were 2,424 DHS sites where three species displayed increased chromatin
431 accessibility relative to rhesus macaque; 794 in human, chimpanzee, and orangutan; 691 in
432 human, gorilla, and orangutan; 939 in chimpanzee, gorilla, and orangutan **(Figure 4C, Table 1)**.
433 There were 2,146 sites where three species displayed decreased accessibility relative to rhesus
434 macaque; 906 in human, chimpanzee, and orangutan; 554 in human, gorilla, and orangutan; 686
435 in chimpanzee, gorilla, and orangutan **(Figure 4D, Table 1)**.

436 ***Comparison to Previous Study with Fewer Species***

437 To test our new method for identifying differences in chromatin accessibility across five
438 species, we compared our results with those from our previous study that used individual
439 pairwise edgeR (Robinson et al. 2010) comparisons for human, chimpanzee, and rhesus macaque
440 (Shibata et al. 2012). We used the same raw fastq files for the human, chimpanzee, and rhesus
441 macaque samples. Due to updates in the analysis pipeline **(Supplementary Table 5)**, not all of
442 the DHS sites that were previously characterized were also identified as DHS sites in this study
443 **(Supplementary Table 5)**.
444 Of the DHS sites that were previously characterized by Shibata et al. as being human-
445 specific increased accessibility, human-specific decreased accessibility, chimpanzee-specific
446 increased accessibility, or chimpanzee-specific decreased accessibility, we find that 88-98% of
447 our more recent set of DHS sites that overlap them displayed a consistent call **(Table 2)**. The
448 additional gorilla and orangutan DNase-seq data in this study allows us to fill in missing branch
449 data and gauge the accuracy of our previous classification of human-specific or chimpanzee-
450 specific changes. For 342 DHS sites that overlap sites previously characterized by Shibata et al.

451 as being human-specific increased accessibility, 245 (72%) are still characterized as human-
452 specific increased accessibility after including gorilla and orangutan, while 91 (27%) are now
453 characterized as increased accessibility in human and at least one other species **(Table 2)**. For
454 234 DHS sites that overlap previously called chimpanzee-specific increased accessibility, 105
455 (45%) are still characterized as chimpanzee-specific increased accessibility after including
456 gorilla and orangutan, while 114 (49%) are now characterized as increased accessibility in
457 chimpanzee and at least one other species **(Table 2)**.

458 A similar trend was detected for previously identified human-specific or chimpanzee-
459 specific decreased accessibility. For 146 DHS sites that overlap sites previously characterized as
460 human-specific decreased accessibility, 21 (14%) are still characterized as human-specific
461 decreased accessibility even after including orangutan and gorilla, while 109 (75%) are now
462 characterized as decreased accessibility in human and at least one other species **(Table 2)**. For 98
463 DHS sites that overlap sites previously called chimpanzee-specific decreased accessibility, 17
464 (17%) are still characterized as chimpanzee-specific decreased accessibility even after including
465 orangutan and gorilla, while 69 (70%) are now characterized as decreased accessibility in
466 chimpanzee and at least one other species **(Table 2)**.

467 For 1,154 DHS sites that overlap sites Shibata et al. (2012) characterized as common
468 between human, chimpanzee, and rhesus macaque, 95% (1,100) displayed a non-differential call
469 among these same species using our pipeline **(Table 2)**. After adding gorilla and orangutan, 90%
470 (1,043) of the DHS sites are still characterized as non-differential, while 5% (57) displayed
471 changes in accessibility on branches not considered by Shibata et al., namely gorilla, orangutan,
472 or both **(Table 2)**.

473 Together, this indicates that adding chromatin accessibility data from additional primate
474 species allows us to identify a substantial subset of DHS sites that have experienced changes in
475 chromatin accessibility across multiple species during evolution.

476 ***DHS Sites with Decreased Accessibility are Enriched for Proximal Elements and DHS Sites***
477 ***with Increased Accessibility are Enriched for Distal Elements***

478 After identifying and classifying DHS sites, we next determined their location in the
479 human genome relative to previously annotated proximal and distal elements. We used the
480 HoneyBadger2 annotations (see Materials and Methods), which are predicted promoters or
481 enhancers based on histone marks identified in human cells and tissues as part of the Roadmap
482 Epigenomics project (Roadmap Epigenomics Consortium et al. 2015). We overlapped these
483 annotations to characterize each DHS site identified in this study as a proximal element, distal
484 element, or unannotated region.

485 For DHS sites that are not differential between primate species, 22% (11,850) of these
486 regions overlap with proximal elements, 57% (30,371) overlap with distal elements, and the
487 remaining 20% (10,857) are unannotated **(Figure 5A, Supplementary Table 6)**. All DHS sites
488 with increased accessibility relative to rhesus macaque display a substantially depleted amount of
489 proximal element overlap compared to the non differential DHS sites (human: 2%; chimpanzee:
490 4%; gorilla: 9%; orangutan: 10%; human-chimpanzee: 3%; human-chimpanzee-gorilla: 5%)
491 **(Figure 5A, Supplementary Table 6)**. Conversely, DHS sites with decreased accessibility
492 relative to rhesus macaque overlap proximal elements to a similar degree as non differential DHS
493 sites (human: 18%; chimpanzee: 21%; gorilla: 11%; orangutan: 20%; human-chimpanzee: 34%;
494 human-chimpanzee-gorilla: 10%) **(Figure 5A, Supplementary Table 6**). These results indicate
495 that decreased accessibility changes are more likely to be associated with proximal elements,

496 while increased accessibility changes are more likely to be associated with distal elements. In
497 every category of accessibility changes, there are substantially more distal than proximal
498 elements, which is consistent with other studies (Schmidt et al. 2010; Villar et al. 2015).
499 We note that all of these proximal and distal annotations are from human tissues, which
500 allows us to make specific inferences about comparisons only to human. There is not yet a
501 similar Roadmap effort for non-human primate species. Categories of accessibility increases that
502 include human (human-specific, human-chimpanzee, and human-chimpanzee-gorilla) have the
503 lowest amount of overlap with unannotated regions of the genome. DHS sites with increased
504 accessibility specific to chimpanzee, gorilla, or orangutan all have much higher overlaps with
505 unannotated regions, with orangutan-specific increased accessibility showing the highest degree
506 of overlap with unannotated regions **(Figure 5A, Supplementary Table 6**). This is expected
507 since orangutan is the most distantly related great ape species in our study. Similarly, we find
508 that DHS sites with decreased accessibility in human (human-specific, human-chimpanzee, and
509 human-chimpanzee-gorilla) have a higher overlap with unannotated regions compared to DHS
510 sites with decreased accessibility specific to chimpanzee, gorilla, and orangutan. This is also
511 expected since DHS sites with decreased accessibility in non-human primates will by definition
512 have higher chromatin accessibility signals in human fibroblasts.

513 ***Evolutionary Changes in Accessibility are Associated with Cell-type Specificity***

514 We calculated cell-type specificity (see Materials and Methods) for the union set of DHS
515 sites detected in primate fibroblasts by comparing them to a much larger set of DHS sites
516 detected in 125 different human cell and tissue types (Thurman et al. 2012). A cell-type
517 specificity score close to 1 indicates the DHS site is present in only a few of the 125 tissues and

518 cell types, while a score near 0 indicates that the DHS site is present in almost all of the 125
519 tissues and cell types.

520 As with the proximal and distal annotations, we can make inferences about evolutionary
521 changes in chromatin accessibility only for DHS sites that overlap the human annotations. The
522 union set of the DHS sites we identified show a continuum of cell-type specificity scores with
523 DHS sites from different human cell types (**Figure 5B**). 1,914 (2%) of our DHS sites overlapped
524 DHS sites found in all 125 tissues and cell types, while 3,873 (4%) of our DHS sites were not
525 found in any of previously tested tissues and cell types.

526 We then analyzed the distribution of cell type specificity scores in distal and proximal
527 DHS sites that displayed changes in chromatin accessibility. In general, distal elements have
528 higher specificity scores than proximal elements **(Figure 5C,D)**, consistent with previous studies
529 (Thurman et al. 2012).

530 For proximal elements showing increases in accessibility, tissue specificity is higher on
531 all four species branches than on the two internal branches (one sided Wilcoxon test comparing
532 pooled distributions of external vs internal; $P = 2.03\text{x}10^{-28}$) **(Figure 5C)**. The opposite pattern is
533 evident for decreases in accessibility (one sided Wilcoxon test comparing pooled distributions of
534 external vs internal; $P = 3.52\text{x}10^{-32}$) **(Figure 5C)**. Since all changes on the internal branches are
535 more ancient than those on external branches, this result hints at the possibility that degree of
536 chromatin accessibility is positively correlated with broader utilization across cell types. One
537 possible explanation is that increases in chromatin accessibility raise the likelihood that a
538 proximal regulatory element is co-opted for use by another tissue. The same trends are observed
539 for distal elements, with the exception of distal sites having higher tissue specificity scores,

540 which is expected since distal chromatin accessible sites are more likely to be cell type specific
541 than proximal elements (**Figure 5D**).
542 For proximal elements showing changes in chromatin state, the human branch shows
543 lower cell type specificity compared to the three other species for increases (one sided Wilcoxon
544 test with Bonferroni correction; $P_{H:C} = 1.62\text{x}10^{-5}$; $P_{H:G} = 1.11\text{x}10^{-6}$; $P_{H:O} = 4.69\text{x}10^{-11}$) and
545 higher cell type specificity for decreases (one sided Wilcoxon test with Bonferroni correction;
546 $P_{H:C} = 0.004$; $P_{H:G} = 0.036$; $P_{H:O} = 0.004$) **(Figure 5C)**. The same pattern is present for distal
547 elements, both for increases and decreases (one sided Wilcoxon test with Bonferroni correction;
548 Increases: $P_{H:C} = 4.55\text{x}10^{-70}$; $P_{H:G} = 1.18\text{x}10^{-44}$; $P_{H:O} = 3.47\text{x}10^{-133}$; Decreases: $P_{H:C} = 1.11\text{x}10^{-6}$;
549 $P_{H:G} = 2.54\text{x}10^{-3}$; $P_{H:O} = 2.38\text{x}10^{-8}$) **(Figure 5D)**. This may reflect an ascertainment bias arising
550 from relying on human tissue comparisons for the cell type specificity score.

551 ***Species-specific Chromatin Changes are Enriched in Positive Selection***

552 To investigate the evolutionary significance of species-specific changes in chromatin
553 accessibility, we tested each DHS site for signatures of positive selection on the human,
554 chimpanzee, and gorilla branches separately (see Materials and Methods). Testing for positive
555 selection required additional filtering of DHS sites (see Materials and Methods), resulting in a
556 reduced set of 87,431 DHS sites used in this analysis. The figure of merit in these analyses is ζ
557 (zeta), the ratio of substitution rates within a DHS site on a given branch relative to the rest of the
558 tree in comparison to that ratio for a collection of proxy neutral sites (Wong and Nielsen 2004;
559 Haygood et al. 2007; Haygood et al. 2010). Similar to the analogous and more familiar ω
560 (omega), high values of ζ indicate positive selection, values near 1 indicate neutrality, and low
561 values indicate negative selection.

562 Putative non functional elements display a relatively tight distribution of ζ on the human
563 branch centered around 1 (**Figure 6A**), confirming they are a good proxy for neutral evolution in
564 non-coding regions of the genome. Non differential DHS sites (those where the degree of
565 chromatin accessibility does not change on any branch) have a distribution of ζ on the human
566 branch that is centered significantly below 1 (one sided Wilcoxon test; $P = 1.57x10^{-283}$) (**Figure**
567 **6A**), consistent with ongoing negative selection. Additionally, the distribution of ζ values is
568 much broader for non differential DHS sites compared to putative non functional sites, with a
569 small fraction showing elevated substitution rates on the human branch that are consistent with
570 positive selection.

571 DHS sites that showed a change in chromatin accessibility on the human branch have
572 positively-shifted distributions of ζ on the human branch relative to non differential DHS sites
573 (**Figure 6A**). This suggests that both increases and decreases are accompanied by enrichment for
574 a combination of relaxed selection and positive selection on the same branch. As expected, this
575 enrichment is less pronounced when the accessibility change occurs on a different branch of the
576 phylogeny: the distributions of ζ on the human branch are higher when the chromatin state
577 change occurred on the human branch rather than the gorilla or orangutan branches, and this is
578 true for both increases and decreases (one sided Wilcoxon test with Bonferroni correction;
579 Increases: $P_{H:C} = 0.12$, $P_{H:G} = 2.17x10^{-10}$, $P_{H:O} = 6.39x10^{-6}$, $P_{H:H-C} = 9.36x10^{-13}$, $P_{H:H-C-G} =$
580 $1.40x10^{-10}$; Decreases: $P_{H:C} = 0.006$, $P_{H:G} =4.13x10^{-5}$, $P_{H:O} = 2.26x10^{-5}$, $P_{H:H-C} = 0.03$, $P_{H:H-C-G} =$
581 $1.92x10^{-4}$), although these differences are all modest in magnitude (**Figure 6A**).

582 For the human-specific accessibility changes, we tested for enrichment of positive
583 selection of these regions in the human branch relative to either the chimpanzee or gorilla
584 branches. We performed a similar comparison for the chimpanzee-specific accessibility changes

585 by testing for enrichment of positive selection on the chimpanzee branch relative to either the
586 human or gorilla branches. Finally, we tested for enrichment of gorilla-specific accessibility
587 changes on the gorilla branch relative to the human or chimpanzee branches (**Figure 6B**). None
588 of the Fisher's exact tests were significant after Bonferroni correction, even though the data
589 trends in the expected patterns (e.g., human-specific changes have more selection on the human
590 branch, etc.).

591 As a control, non differential DHS sites show no significant differences in positive
592 selection between branches (**Figure 6B**). Additionally, putatively non functional sites in the
593 genome do not display an enrichment of positive selection (**Figure 6B**). These results suggest
594 that evolutionary changes in chromatin accessibility between species are phylogenetically
595 correlated with an enrichment of positive selection.

596 Next, we investigated the converse: whether signatures of positive selection on individual
597 regulatory elements are generally limited to branches where the state change occurred. This is
598 clearly not the case: DHS sites with increased accessibility on the human branch show positive
599 selection on the chimpanzee branch almost as often as on the human branch (**Figure 6C**). The
600 same pattern is evident for increased accessibility on the chimpanzee branch and for decreased
601 accessibility on the human or chimpanzee branch (**Figure 6C**). These results suggest that
602 positive selection often acts on the DNA sequence of DHS sites in ways that do not affect
603 chromatin state. Interestingly, some DHS sites show evidence of positive selection on both
604 branches (**Figure 6C**). Given that 1.96% of DHS sites contain signatures of positive selection in
605 the human branch, this result suggests that positive selection is not distributed randomly across
606 the regulatory genome, but occurs repeatedly in a very small percentage of sites (0.12%).

607 On average, 5.4% of DHS sites that show species-specific accessibility changes for human,
608 chimpanzee, and gorilla are highly constrained in sequence evolution, with nucleotide
609 substitution rates that are significantly lower than the neutral expectation across vertebrates
610 (**Figure 6D**). In contrast, DHS sites that do not change state are enriched for highly constrained
611 sites in comparison to species-specific changes for human, chimpanzee, and gorilla (Fisher's
612 exact test, one-sided with Bonferroni adjust for 3 comparisons; $P_{ND:H} = 9.10x10^{-6}$;
613 $P_{ND:C}=2.42x10^{-6}$; $P_{ND:G}=3.18x10^{-4}$) (**Figure 6D**). Together, these results suggest that positive
614 selection contributes to species-specific chromatin accessibility increases and decreases, while
615 purifying selection contributes to the conservation of non differential DHS sites.

616 **Discussion**

617 We developed a new method to identify regions of differential chromatin accessibility
618 among multiple species, based on a negative binomial generalized linear model. This method
619 does not rely on thresholding and is therefore able to detect subtle differences in degree of
620 chromatin accessibility that are obscured using conventional approaches. Furthermore, unlike
621 methods based on pairwise comparison, our method is readily scalable to multiple taxa. Here we
622 applied it to DNase-seq data from cultured skin fibroblasts obtained from five species. While the
623 majority of DHS sites were not quantitatively distinct between species, we identified over 35,000
624 DHS sites with significant differences in chromatin accessibility between human, chimpanzee,
625 gorilla, orangutan, and rhesus macaque. Of those, approximately 65% are likely the result of a
626 single evolutionary change in chromatin state that occured on either an internal or external
627 branch, while the remainder imply multiple evolutionary changes in state.

628 Our results are largely congruent with our earlier study (Shibata et al. 2012) that used a
629 threshold-based multiple pairwise comparison approach and considered three primate species

630 (human, chimpanzee, and rhesus macaque). Here, the use of five species provides additional
631 confidence in the identification of species-specific accessibility changes and also allows for the
632 identification of changes that likely occurred multiple times throughout evolution. For these
633 multiple events, the method we developed does not characterize how exactly they occurred. That
634 characterization will be the subject of future work using a likelihood analysis that incorporates
635 the phylogenetic information and models evolutionary processes (Felsenstein 1973; Hansen
636 1997; Felsenstein 2008; Paradis and Schliep 2019).

637 As mentioned in the results, we identified substantially more accessibility increases than
638 decreases. It seems in principle unlikely that increases and decreases in accessibility actually
639 occur at such different rates, since, if true, primate genomes would eventually become saturated
640 with open chromatin regions. The same asymmetry was observed previously by us (Shibata et al.
641 2012) and other groups (Villar et al. 2015; Reilly et al. 2015; Emera et al. 2016) using
642 conventional pairwise comparisons and thresholding, so the source is unlikely to lie in the
643 method used to identify chromatin state changes. Instead, it seems likely that the asymmetry is
644 an ascertainment bias that derives, in part, from unequal statistical power to call gains and losses,
645 though the exact basis of the bias is currently unclear.

646 Finding so many DHS site differences in non-human primates is a fascinating result with
647 implications for understanding the evolution of transcriptional regulation. Nevertheless, we also
648 suggest that these results describing cell type specificity should be interpreted carefully. One
649 non-biological possible scenario for such enrichment is that there is ascertainment bias in our
650 analyses due to the cell-type specificity score being based entirely on data from human, a
651 limitation imposed by lack of relevant data from other primate species. Additionally, many of
652 the chromatin accessibility datasets analyzed by Thurman et al. that we used to calculate the cell-

653 type specificity scores are from fibroblast-like cells. Some of the DHS sites that we identified as
654 being used in many cell types may in fact be used only in fibroblast-like tissues. Although the
655 patterns of positive selection that we detected are consistent with expectations, none of the tests
656 found statistically significant enrichment on the human, chimpanzee, and gorilla branches. This
657 may be due to our method of positive selection detection relying on human functional
658 annotations to identify proxy neutral regions, which may result in a loss of power with increasing
659 phylogenetic distance.

660 Interestingly, our results suggest that DHS sites are not homogenous from either a
661 functional or an evolutionary perspective. Those near transcription start sites (including likely
662 core promoter regions) differ from DHS sites that are distant (including classic enhancers and
663 other kinds of distal elements) in several regards. DHS sites that show no state changes across
664 the species surveyed are enriched in conserved nucleotides, consistent with greater functional
665 significance. Compared with proximal DHS sites, gains in chromatin accessibility in distal sites
666 are more likely to show same branch-specific signatures of positive selection, as might be
667 expected if these DHS sites are contributing to changes in gene regulation. These and other
668 trends we observed suggest that functional constraints and opportunities differ markedly among
669 classes of DHS sites. Additional studies will be needed to delineate these distinct classes of
670 likely regulatory elements and to understand how evolutionary mechanisms operate on their
671 chromatin state and underlying DNA sequence.

672 Functional characterization studies will be necessary to understand these regions and
673 their contribution to species-specific gene expression patterns and organismal traits. High-
674 throughput reporter assays such as MPRA (Klein et al. 2018) and population STARR-seq
675 (Vockley et al. 2015) can quantify the impact of these differentially utilized regulatory regions,

676 as well as variants within these regions. In addition, methods such as CRISPR (Diao et al. 2017;
677 Klann et al. 2017) can characterize the impact of these regions in their natural context, including
678 identifying the correct target gene(s) for these regulatory elements. Finally, additional replicates
679 from these species can provide characterization of variability within each species. While
680 obtaining data from additional tissues for primate species is not possible for most tissues,
681 generation of induced pluripotent stem cells (iPSCs) followed by differentiation (Gallego
682 Romero et al. 2018), will provide insights into how these differential chromatin signals translate
683 into different cell types across many species.

684 While we used our negative binomial generalized linear model method to identify
685 differences in chromatin accessibility between five species, we believe this strategy can be used
686 for quantitative comparisons across many tissues, cell types, and time-series experiments. In
687 addition to DNase-seq, we expect this method can be readily applied to any count-based data
688 type such as RNA-seq, ATAC-seq, and ChIP-seq. We also note that this method is scalable, as
689 increasing the number of groups or replicates requires simply changing the design matrix and
690 contrasts.

691 **Acknowledgments**

692 We would like to thank Terry Gaasterland for her help developing the tiered mapping
693 approach. We thank the Duke Sequencing and Genomic Technologies Shared Resource for
694 sequencing. This work was supported by the National Science Foundation [HOMINID 0827552
695 to G.A.W]; the National Institute of Mental Health at the National Institutes of Health
696 [5R01MH105472 to G.E.C and G.A.W]; and a generous donation from Dr. Howard Clark.
697

698 **Figure and table legends**

699 **FIG. 1.—**Classification of DHS sites. (A-D) Density plots showing the beta values for human
700 (black), chimpanzee (green), gorilla (blue), and orangutan (orange). Beta values near zero
701 indicate the accessibility for that species is similar to rhesus macaque. Positive beta values
702 indicate the accessibility for that species is increased relative to rhesus macaque. Negative beta
703 values indicate the accessibility for that species is decreased relative to rhesus macaque. (A) Non
704 differential sites. (B) Human-specific increases in accessibility. (C) Human-specific decreases in
705 accessibility. (D) Accessibility increases in human and gorilla. (E) Pie chart showing the number
706 and proportion of DHS sites that 1) are non differential; 2) have accessibility changes likely due
707 to a single event; and 3) have accessibility changes that are due to multiple events. Percentages
708 are of the total number of DHS sites. (F) Pie chart showing the number and proportion of each
709 type of single event. The slice labeled "H-C-G-O and Rhesus macaque" contains changes that
710 happened on either the human-chimpanzee-gorilla-orangutan inner branch or the rhesus macaque
711 species branch. Percentages are of the number of DHS sites with a single change. (G) Bar chart
712 showing the relative proportions of increases and decreases in accessibility. Numbers at the top
713 are the total number of DHS sites in each category. Numbers in or just above the orange bar are
714 the number of DHS sites with decreased accessibility. Numbers at the bottom of the green bar
715 are the number of DHS sites with increased accessibility. H-C: Human-chimpanzee internal
716 branch. H-C-G: Human-chimpanzee-gorilla internal branch. H-C-G-O: Human-chimpanzee-
717 gorilla-orangutan internal branch.

718

719 **FIG. 2.—**Species-specific changes in chromatin accessibility. Phylogenetic tree with divergence
720 dates (to scale). UCSC Genome Browser screenshots of representative DHS sites for species-

721 specific DHS sites for (A) increased accessibility and (B) decreased accessibility. Heatmaps of
722 normalized scores are rank-ordered DHS sites based on hierarchical clustering.
723
724 **FIG. 3. —**Internal branch changes in chromatin accessibility. Phylogenetic tree with divergence
725 dates (to scale). UCSC Genome Browser screenshot of a representative DHS site changes that
726 likely occurred before the human-chimpanzee split (top) and human-chimpanzee-gorilla split
727 (bottom). Heatmaps of normalized scores are rank-ordered DHS sites based on hierarchical
728 clustering.
729
730 **FIG. 4. —**Changes in chromatin accessibility due to multiple evolutionary events.
731 Heatmaps of normalized scores that are rank-ordered DHS sites based on hierarchical clustering.
732 **(**A) Two species have increased chromatin accessibility relative to rhesus macaque. (B) Two
733 species have decreased chromatin accessibility relative to rhesus macaque. (C) Three species
734 have increased chromatin accessibility relative to rhesus macaque. (D) Three species have
735 decreased chromatin accessibility relative to rhesus macaque.
736
737 **FIG. 5. —**Chromatin accessibility changes relative to proximal/distal location and cell-type
738 specificity. (A) The percentage of proximal elements, distal elements, and unannotated elements
739 for each category of DHS sites. (B) Histogram of specificity scores for DHS sites identified in
740 this study compared to DHS sites detected in 125 different cell types (Thurman et al). A high
741 specificity score indicates the DHS site is specific to a small number of cell types. A low
742 specificity score indicates the DHS site is shared across many cell types. The DHS site categories

743 are separated into proximal elements (C) and distal elements (D). H-C: human-chimpanzee
744 internal branch. H-C-G: human-chimpanzee-gorilla internal branch.
745
746 **FIG. 6. —**Effect of natural selection on increases and decreases in chromatin accessibility.
747 (A) Distribution of the ratio of evolution ζ (zeta) in the human branch for DHS sites. The dashed
748 green line depicts the critical value where the human zeta value becomes significant ($P < 0.05$).
749 Zeta values around 1 are expected to be neutral and below 1 are expected to be constrained. (B)
750 Percentages of DHS sites that are significant for positive selection ($P < 0.05$). Each DHS site was
751 tested with 3 different foregrounds: human, chimpanzee, and gorilla. (C) Scatterplots of zeta
752 values for DHS sites with significant positive selection on the human branch (purple diamond),
753 the chimpanzee branch (blue cross), or both the human and chimpanzee branches (black solid
754 circle). Zeta values for the human branch are on the x-axis and zeta values for the chimpanzee
755 branch are on the y-axis. The kernel density depicts non functional sites. (Top left) increased
756 accessibility in human. (Top right) decreased accessibility in human. (Bottom left) increased
757 accessibility in chimpanzee. (Bottom right) decreased accessibility in chimpanzee. (D)
758 Percentages of DHS sites that are highly constrained (median vertebrate PhastCons > 0.9). H-C:
759 human-chimpanzee internal branch. H-C-G: human-chimpanzee-gorilla internal branch. ***: p-
760 value < 0.001. **: p-value < 0.01.
761
762 **Table 1**
763 Number of DHS sites in each category
764
765 **Table 2**

766 Comparison to earlier study (Shibata et al., 2012, PLoS Genetics)

767 NOTE. —Increase refers to an increase in chromatin accessibility compared to rhesus macaque.

768 Decrease refers to a decrease in chromatin accessibility compared to rhesus macaque.

769

770 **FIG. S1. —**Average number of per-species DHS sites at each level of subsampling. There is a

771 large range in the number of mapped reads between the samples in this study (Supplementary

772 Table 2). We performed peak calling on random subsets of varying sizes (10 million, 15 million,

773 20 million, 25 million) for each sample. Trajectories are shown.

774

775 **FIG. S2. —**Distribution of beta values for chromatin accessibility changes due to a single

776 evolutionary event. Density plots showing the beta values for human (black), chimpanzee

777 (green), gorilla (blue), and orangutan (orange). Beta values near zero indicate the accessibility

778 for that species is similar to rhesus macaque. Positive beta values indicate the accessibility for

779 that species is increased relative to rhesus macaque. Negative beta values indicate the

780 accessibility for that species is decreased relative to rhesus macaque. Each plot is for a different

781 category of differential site.

782

783 **FIG. S3. —**Distribution of beta values for chromatin accessibility changes due to multiple

784 evolutionary events. Density plots showing the beta values for human (black), chimpanzee

785 (green), gorilla (blue), and orangutan (orange). Beta values near zero indicate the accessibility

786 for that species is similar to rhesus macaque. Positive beta values indicate the accessibility for

787 that species is increased relative to rhesus macaque. Negative beta values indicate the

788 accessibility for that species is decreased relative to rhesus macaque. Each plot is for a different
789 category of differential site.
790
791 **FIG. S4. —**Principal components analysis of the samples analyzed in this study with (right side)
792 and without (left side) three additional chimpanzee samples from Pizzollo et al. 2018. (Top)
793 Principal component 1 (x-axis) vs. principal component 2 (y-axis). (Middle) Principal
794 component 1 (x-axis) vs. principal component 3 (y-axis). (Bottom) Principal component 1 (x-
795 axis) vs. principal component 3 (y-axis). The original samples are represented by filled circles
796 and the Pizzollo et al. samples are represented by filled diamonds. Human samples are in black,
797 chimpanzee samples are in green, gorilla samples are in blue, orangutan samples are in orange,
798 and rhesus macaque samples are in purple.
799
800 **Table S1**
801 Sample information
802
803 **Table S2**
804 Mapping and LiftOver statistics
805 NOTE. —(a) Human replicates were not lifted over; the numbers are repeated for consistency.
806 Percentages of "After lift over" reads are of mapped reads. (b) Final counts are after reads
807 mapped to sex chromosomes are removed.
808
809 **Table S3**
810 DHS site counts at different steps in the analysis pipeline

811

812 **Table S4**

813 Constraint matrices

814

815 **Table S5**

816 Comparison to earlier study (Shibata et al., 2012, PLoS Genetics)

817

818 **Table S6**

819 Location of DHS sites

820 NOTE. —Percentages are of the total number of DHS sites. H-C: human-chimpanzee internal

821 branch. H-C-G: human-chimpanzee-gorilla internal branch.

822 **References**

823 Afgan E, et al. 2018. The Galaxy platform for accessible, reproducible and collaborative

824 biomedical analyses: 2018 update. Nucleic Acids Res. 46(W1):W537-W544.

825 Barbosa-Morais NL, et al. 2012. The evolutionary landscape of alternative splicing in vertebrate

826 species. Science 338(6114):1587-1593.

827 Berrio A, Haygood R, Wray GA. Identifying branch-specific positive selection throughout the

828 regulatory genome using an appropriate neutral proxy. Under review at Genome Biology and

829 Evolution.

830 Blanchette M, et al. 2004. Aligning multiple genomic sequences with the threaded blockset

831 aligner. Genome Res. 14(4):708-715.

832 Blekhman R, Oshlack A, Chabot AE, Smyth GK, Gilad Y. 2008. Gene regulation in primates

833 evolves under tissue-specific selection pressures. PLoS Genet. 4(11):e1000271.

834 Blekhman R, Marioni JC, Zumbo P, Stephens M, Gilad Y. 2010. Sex-specific and lineage-
835 specific alternative splicing in primates. Genome Res. 20(2):180-189.
836 Boyd JL, et al. 2015. Human-Chimpanzee Differences in a FZD8 Enhancer Alter Cell-Cycle
837 Dynamics in the Developing Neocortex. Curr Biol. 25(6):772-779
838 Boyle AP, et al. 2008. High-resolution mapping and characterization of open chromatin across
839 the genome. Cell 132(2):311-322.
840 Brawand D, et al. 2011. The evolution of gene expression levels in mammalian organs. Nature
841 478(7369):343-348.
842 Cáceres M, et al. 2003. Elevated gene expression levels distinguish human from non-human
843 primate brains. Proc Natl Acad Sci U S A. 100(22):13030-13035.
844 Chimpanzee Sequencing and Analysis Consortium 2005. Initial sequence of the chimpanzee
845 genome and comparison with the human genome. Nature 437(7055):69-87.
846 Dermitzakis ET, Clark. AG 2002. Evolution of transcription factor binding sites in Mammalian
847 gene regulatory regions: conservation and turnover. Mol Biol Evol. 19(7):1114-1121.
848 Emera D, Yin J, Reilly SK, Gockley J, Noonan JP. 2016. Origin and evolution of developmental
849 enhancers in the mammalian neocortex. Proc Natl Acad Sci U S A. 113(19):E2617-2626.
850 ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human
851 genome. Nature 489(7414):57-74.
852 Felsenstein J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous
853 characters. Am J Hum Genet. 25(5):471-492.
854 Felsenstein J. 2008. Comparative methods with sampling error and within-species variation:
855 contrasts revisited and revised. Am Nat. 171(6):713-725.

856 Gilad Y, Oshlack A, Smyth GK, Speed TP, White KP. 2006. Expression profiling in primates
857 reveals a rapid evolution of human transcription factors. Nature 440(7081):242-245.
858 Hansen TF. 1997. STABILIZING SELECTION AND THE COMPARATIVE ANALYSIS OF
859 ADAPTATION. Evolution 51(5):1341-1351.
860 Haygood R, Fedrigo O, Hanson B, Yokoyama KD, Wray GA. 2007. Promoter regions of many
861 neural- and nutrition-related genes have experienced positive selection during human evolution.
862 Nat Genet. 39(9):1140-1144.
863 Haygood R, Babbitt CC, Fedrigo O, Wray GA. 2010. Contrasts between adaptive coding and
864 noncoding changes during human evolution. Proc Natl Acad Sci U S A. 107(17):7853-7857.
865 Hernando-Herraez I, et al. 2013. Dynamics of DNA methylation in recent human and great ape
866 evolution. PLoS Genet. 9(9):e1003763.
867 Hinrichs AS, et al. 2006. The UCSC Genome Browser Database: update 2006. Nucleic Acids
868 Res. 34(Database issue):D590-598.
869 Hubisz MJ, Pollard KS, Siepel A. 2011. PHAST and RPHAST: phylogenetic analysis with
870 space/time models. Brief Bioinformatics 12(1):41-51.
871 King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. Science
872 188(4184):107-116.
873 Klann TS, et al. 2017. CRISPR-Cas9 epigenome editing enables high-throughput screening for
874 functional regulatory elements in the human genome. Nat Biotechnol. 35(6):561-568.
875 Klein JC, Keith A, Agarwal V, Durham T, Shendure J. 2018. Functional characterization of
876 enhancer evolution in the primate lineage. Genome Biol. 19(1):99.
877 Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. Nature
878 409(6822):860-921.

879 Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment
880 of short DNA sequences to the human genome. Genome Biol. 10(3):R25.
881 Li H, et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics
882 25(16):2078-2079.
883 Locke DP, et al. 2011. Comparative and demographic analysis of orang-utan genomes. Nature
884 469(7331):529-533.
885 Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for
886 RNA-seq data with DESeq2. Genome Biol. 15(12):550.
887 Neph S, et al. 2012. BEDOPS: high-performance genomic feature operations. Bioinformatics
888 28(14):1919-1920.
889 Odom DT, et al. 2007. Tissue-specific transcriptional regulation has diverged significantly
890 between human and mouse. Nat Genet. 39(6):730-732.
891 Pai AA, Bell JT, Marioni JC, Pritchard JK, Gilad Y. 2011. A genome-wide study of DNA
892 methylation patterns and gene expression levels in multiple human and chimpanzee tissues.
893 PLoS Genet. 7(2):e1001316.
894 Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary
895 analyses in R. Bioinformatics 35(3):526-528.
896 Pizzollo J, et al. 2018. Comparative Serum Challenges Show Divergent Patterns of Gene
897 Expression and Open Chromatin in Human and Chimpanzee. Genome Biol Evol. 10(3):826-839.
898 Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution
899 rates on mammalian phylogenies. Genome Res. 20(1):110-121.
900 Pond SLK, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies.
901 Bioinformatics 21(5):676-679.

902 Prabhakar S, et al. 2008. Human-specific gain of function in a developmental enhancer. Science
903 321(5894):1346-1350.
904 Prescott SL, et al. 2015. Enhancer Divergence and cis-Regulatory Evolution in the Human and
905 Chimp Neural Crest. Cell 163(1):68-83.
906 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic
907 features. Bioinformatics 26(6):841-842.
908 Reilly SK, et al. 2015. Evolutionary changes in promoter and enhancer activity during human
909 corticogenesis. Science 347(6226):1155-1159.
910 Roadmap Epigenomics Consortium, et al. 2015. Integrative analysis of 111 reference human
911 epigenomes. Nature 518(7539):317-330.
912 Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential
913 expression analysis of digital gene expression data. Bioinformatics 26:(1)139-140.
914 Gallego Romero I, Gopalakrishnan S, Gilad Y. 2018. Widespread conservation of chromatin
915 accessibility patterns and transcription factor binding in human and chimpanzee induced
916 pluripotent stem cells. bioRxiv: 466631. doi: https://doi.org/10.1101/466631
917 Rosenbloom KR, et al. 2013. ENCODE data in the UCSC Genome Browser: year 5 update.
918 Nucleic Acids Res. 41(Database issue):D56-63.
919 Scally A, et al. 2012. Insights into hominid evolution from the gorilla genome sequence. Nature
920 483(7388):169-175.
921 Schmidt D, et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of
922 transcription factor binding. Science 328(5981):1036-1040.
923 Schrago CG, Voloch CM. 2013. The precision of the hominid timescale estimated by relaxed
924 clock methods. J Evol Biol 26(4):746-755.

925 Shibata Y, et al. 2012. Extensive evolutionary changes in regulatory element activity during
926 human origins are associated with altered gene expression and positive selection. PLoS Genet.
927 8(6):e1002789.
928 Siepel A, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast
929 genomes. Genome Res. 15(8):1034-1050.
930 Thurman RE, et al. 2012. The accessible chromatin landscape of the human genome. Nature
931 489(7414):75-82.
932 Villar D, et al. 2015. Enhancer Evolution across 20 Mammalian Species. Cell 160(3):554-566.
933 Vockley CM, et al. 2015. Massively parallel quantification of the regulatory effects of noncoding
934 genetic variation in a human cohort. Genome Res. 25(8):1206-1214.
935 Wittkopp PJ, Kalay G. 2011. Cis-regulatory elements: molecular mechanisms and evolutionary
936 processes underlying divergence. Nat Rev Genet. 13(1):59-69.
937 Wong WSW, Nielsen R. 2004. Detecting selection in noncoding regions of nucleotide
938 sequences. Genetics 167(2):949-958.
939 Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet.
940 8(3):206-216.
941 Wu H, Wang C, Wu Z. 2013. A new shrinkage estimator for dispersion improves differential
942 expression detection in RNA-seq data. Biostatistics 14(2):232-243.
943 Yan G, et al. 2011. Genome sequencing and comparison of two nonhuman primate animal
944 models, the cynomolgus and Chinese rhesus macaques. Nat Biotechnol. 29(11):1019-1023.
945 Yu G, Smith DK, Zhu H, Guan Y Lam TT-Y. 2016. ggtree: an R package for visualization and
946 annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol
947 Evol 8(1):28-36.

948 Zeng J, et al. 2012. Divergent whole-genome methylation maps of human and chimpanzee brains
949 reveal epigenetic basis of human regulatory evolution. Am J Hum Genet. 91(3):455-465.
950 Zhang Y, et al. 2008. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 9(9):R137.
951 Zhou X, et al. 2014. Epigenetic modifications are associated with inter-species gene expression
952 variation in primates. Genome Biol. 15(12):547.

FIG. 1

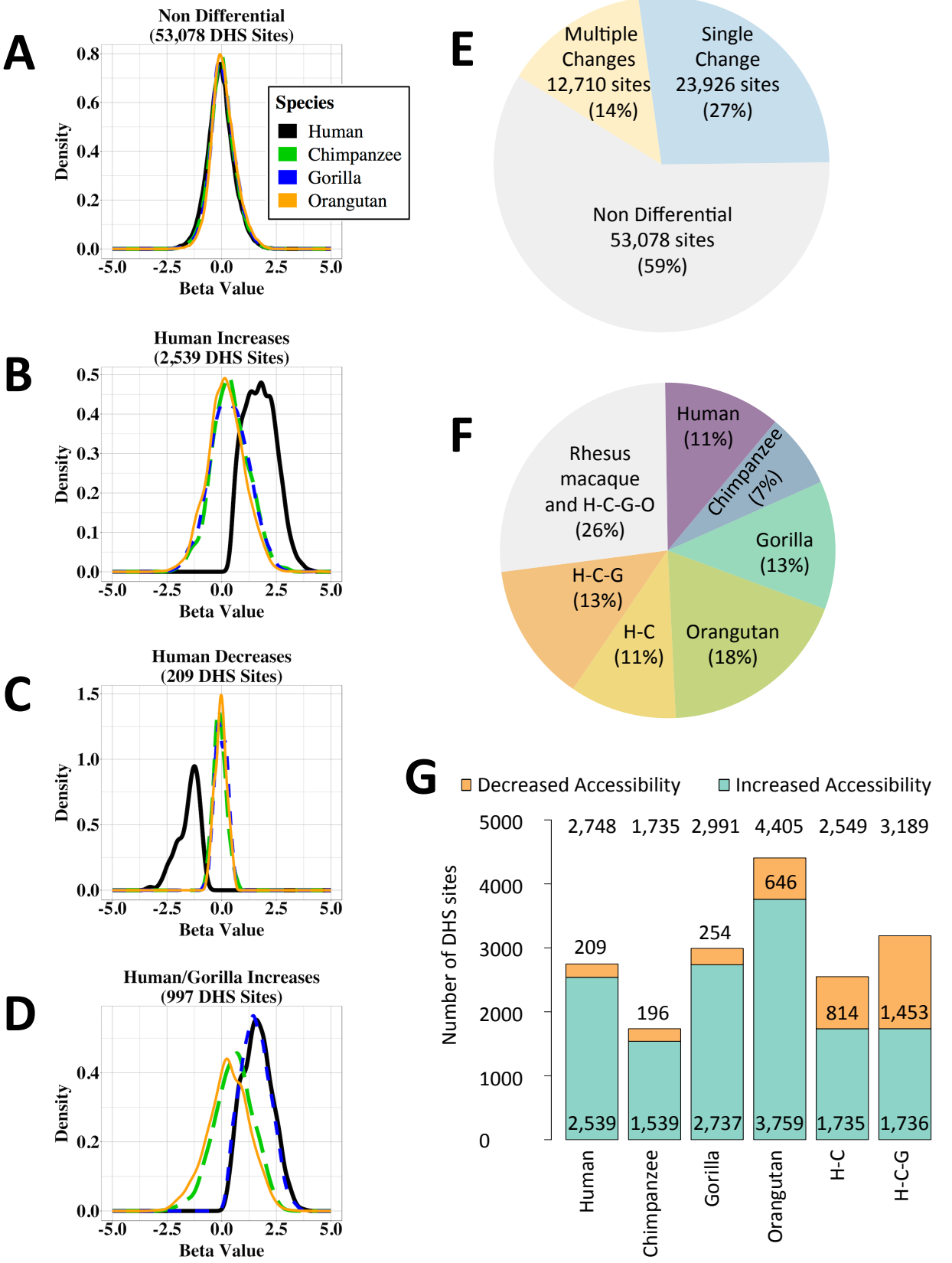
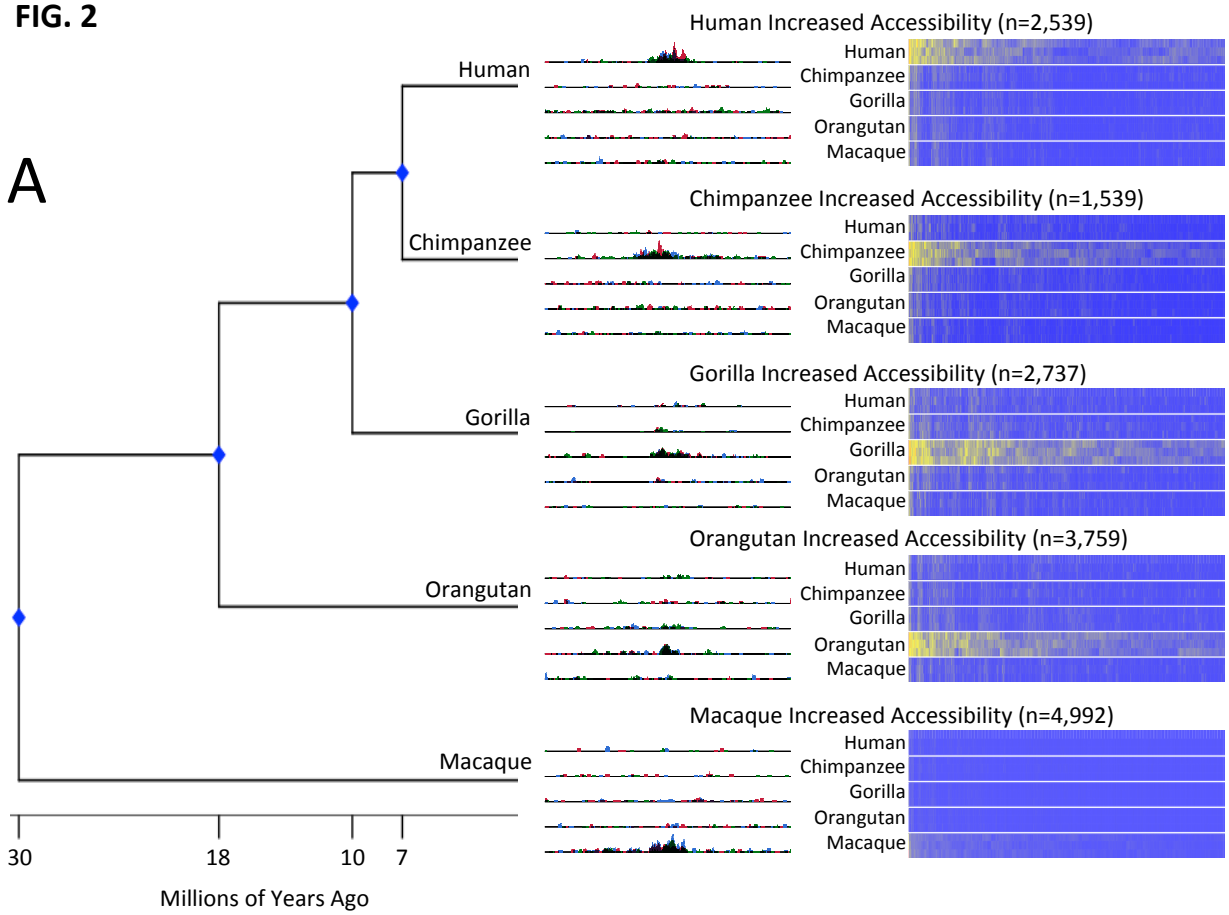

A
Non Differential
(53,078 DHS Sites)
Density
Beta Value
Species
Human
Chimpanzee
Gorilla
Orangutan
B
Human Increases
(2,539 DHS Sites)
C
Human Decreases
(209 DHS Sites)
D
Human/Gorilla Increases
(997 DHS Sites)
E
Multiple Changes 12,710 sites (14%)
Single Change 23,926 sites (27%)
Non Differential 53,078 sites (59%)
F
Human (11%)
Chimpanzee (7%)
Gorilla (13%)
Orangutan (18%)
H-C (11%)
H-C-G (13%)
Rhesus macaque and H-C-G-O (26%)
G
Decreased Accessibility
Increased Accessibility
Number of DHS sites
2,748
1,735
2,991
4,405
2,549
3,189
209
196
254
646
814
1,453
2,539
1,539
2,737
3,759
1,735
1,736
Human
Chimpanzee
Gorilla
Orangutan
H-C
H-C-G

FIG. 2


A
Human
Chimpanzee
Gorilla
Orangutan
Macaque
30
18
10
7
Millions of Years Ago
Human Increased Accessibility (n=2,539)
Chimpanzee Increased Accessibility (n=1,539)
Gorilla Increased Accessibility (n=2,737)
Orangutan Increased Accessibility (n=3,759)
Macaque Increased Accessibility (n=4,992)
Human
Chimpanzee
Gorilla
Orangutan
Macaque

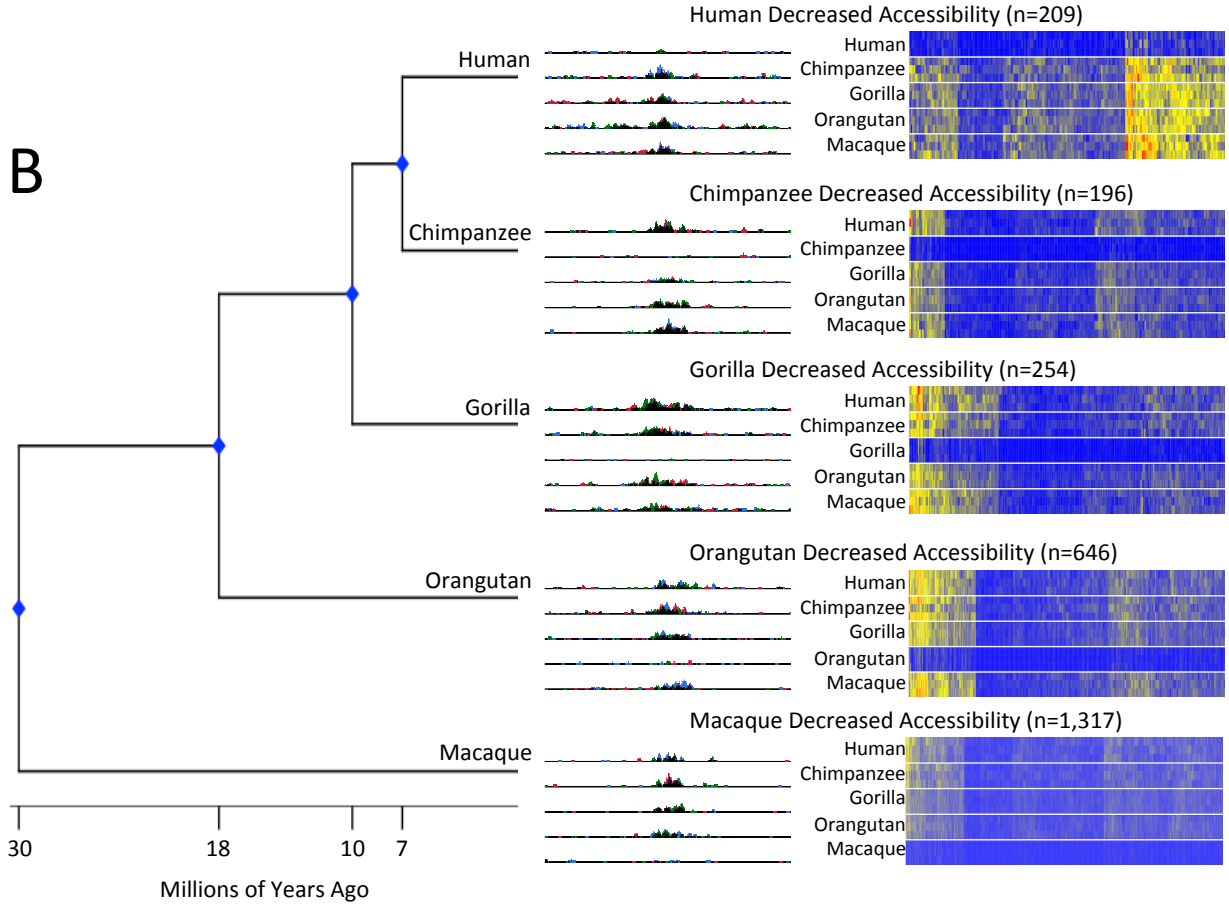

B
Human
Chimpanzee
Gorilla
Orangutan
Macaque
30
18
10
7
Millions of Years Ago
Human Decreased Accessibility (n=209)
Chimpanzee Decreased Accessibility (n=196)
Gorilla Decreased Accessibility (n=254)
Orangutan Decreased Accessibility (n=646)
Macaque Decreased Accessibility (n=1,317)
Human
Chimpanzee
Gorilla
Orangutan
Macaque

FIG. 3

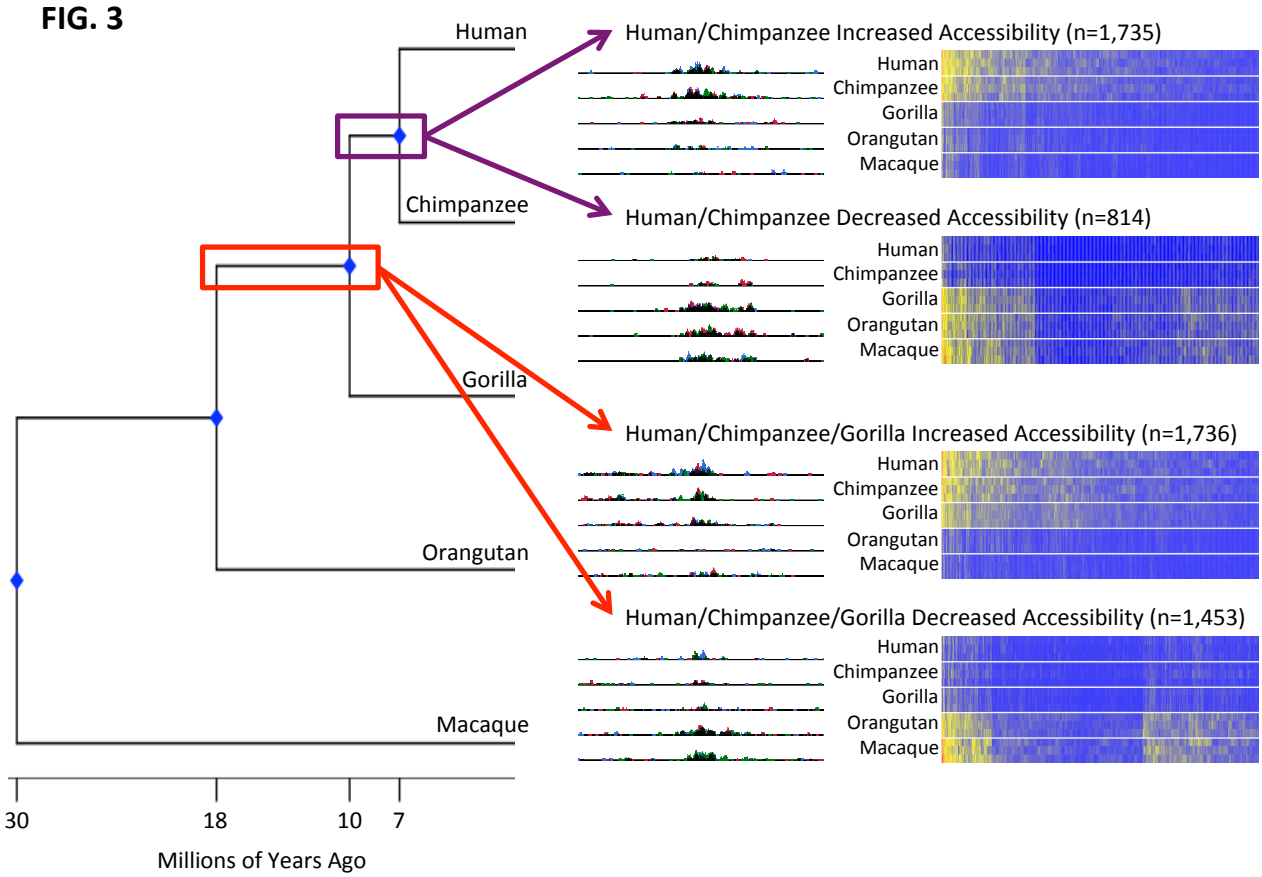

Human
Chimpanzee
Gorilla
Orangutan
Macaque
30
18
10
7
Millions of Years Ago
Human/Chimpanzee Increased Accessibility (n=1,735)
Human/Chimpanzee Decreased Accessibility (n=814)
Human/Chimpanzee/Gorilla Increased Accessibility (n=1,736)
Human/Chimpanzee/Gorilla Decreased Accessibility (n=1,453)
Human
Chimpanzee
Gorilla
Orangutan
Macaque

FIG. 4

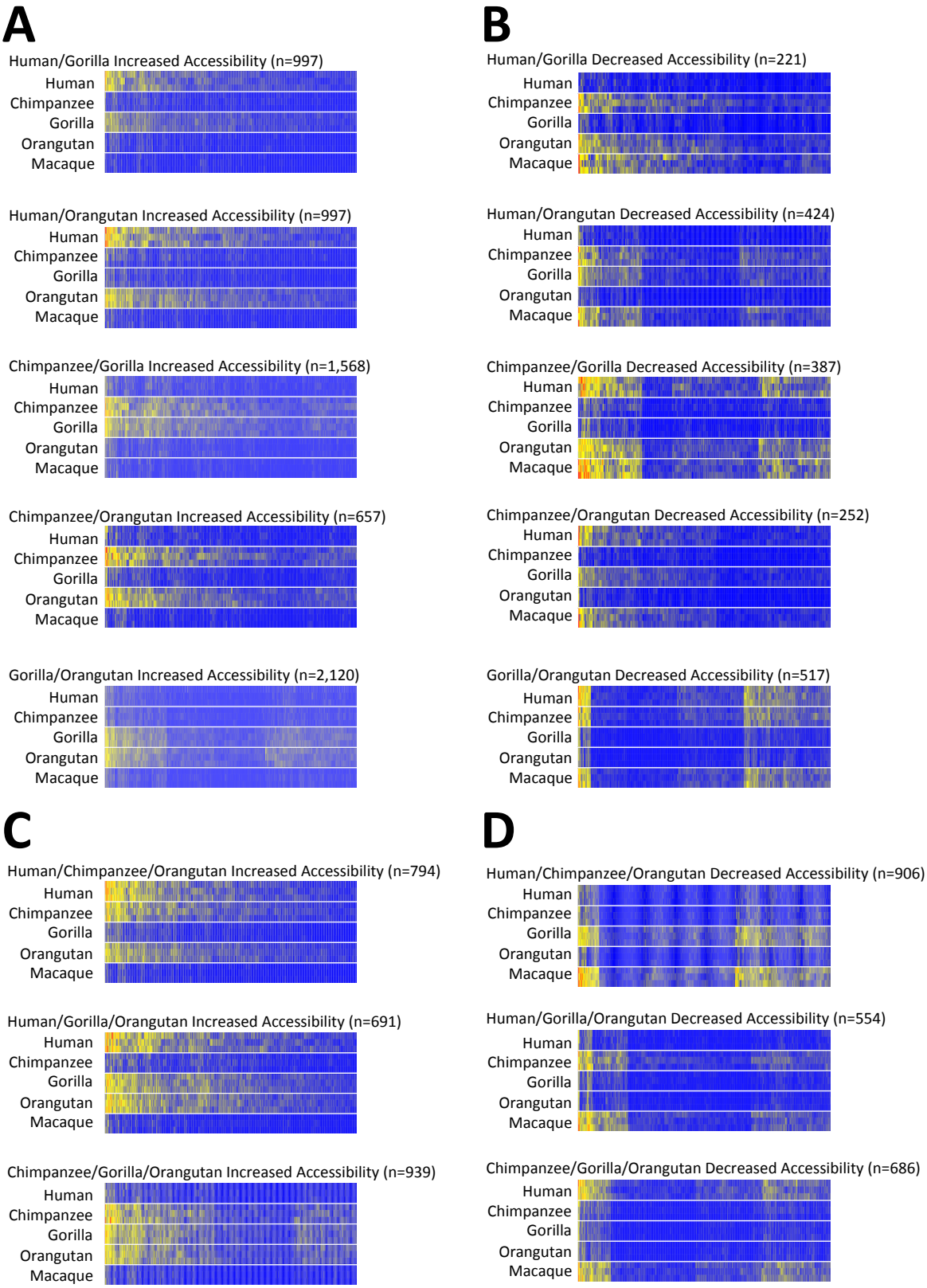

A
Human/Gorilla Increased Accessibility (n=997)
Human/Orangutan Increased Accessibility (n=997)
Chimpanzee/Gorilla Increased Accessibility (n=1,568)
Chimpanzee/Orangutan Increased Accessibility (n=657)
Gorilla/Orangutan Increased Accessibility (n=2,120)
B
Human/Gorilla Decreased Accessibility (n=221)
Human/Orangutan Decreased Accessibility (n=424)
Chimpanzee/Gorilla Decreased Accessibility (n=387)
Chimpanzee/Orangutan Decreased Accessibility (n=252)
Gorilla/Orangutan Decreased Accessibility (n=517)
Human
Chimpanzee
Gorilla
Orangutan
Macaque
C
Human/Chimpanzee/Orangutan Increased Accessibility (n=794)
Human/Gorilla/Orangutan Increased Accessibility (n=691)
Chimpanzee/Gorilla/Orangutan Increased Accessibility (n=939)
D
Human/Chimpanzee/Orangutan Decreased Accessibility (n=906)
Human/Gorilla/Orangutan Decreased Accessibility (n=554)
Chimpanzee/Gorilla/Orangutan Decreased Accessibility (n=686)

FIG. 5

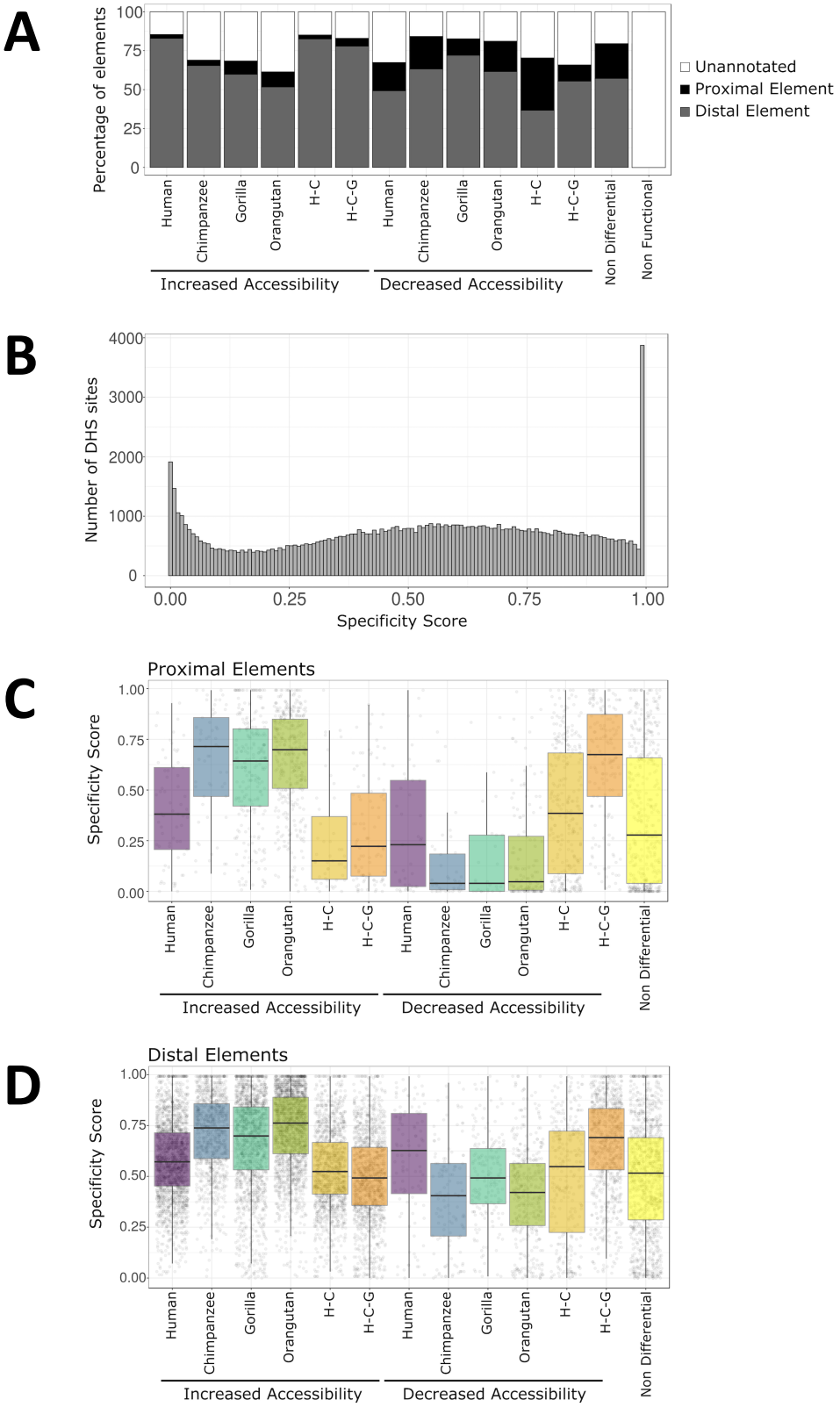

A
Percentage of elements
100
75
50
25
0
Unannotated
Proximal Element
Distal Element
Human
Chimpanzee
Gorilla
Orangutan
H-C
H-C-G
Human
Chimpanzee
Gorilla
Orangutan
H-C
H-C-G
Non Differential
Non Functional
Increased Accessibility
Decreased Accessibility
B
Number of DHS sites
4000
3000
2000
1000
0
0.00
0.25
0.50
0.75
1.00
Specificity Score
C
Proximal Elements
Specificity Score
1.00
0.75
0.50
0.25
0.00
Human
Chimpanzee
Gorilla
Orangutan
H-C
H-C-G
Human
Chimpanzee
Gorilla
Orangutan
H-C
H-C-G
Non Differential
Increased Accessibility
Decreased Accessibility
D
Distal Elements
Specificity Score
1.00
0.75
0.50
0.25
0.00
Human
Chimpanzee
Gorilla
Orangutan
H-C
H-C-G
Human
Chimpanzee
Gorilla
Orangutan
H-C
H-C-G
Non Differential
Increased Accessibility
Decreased Accessibility

FIG. 6

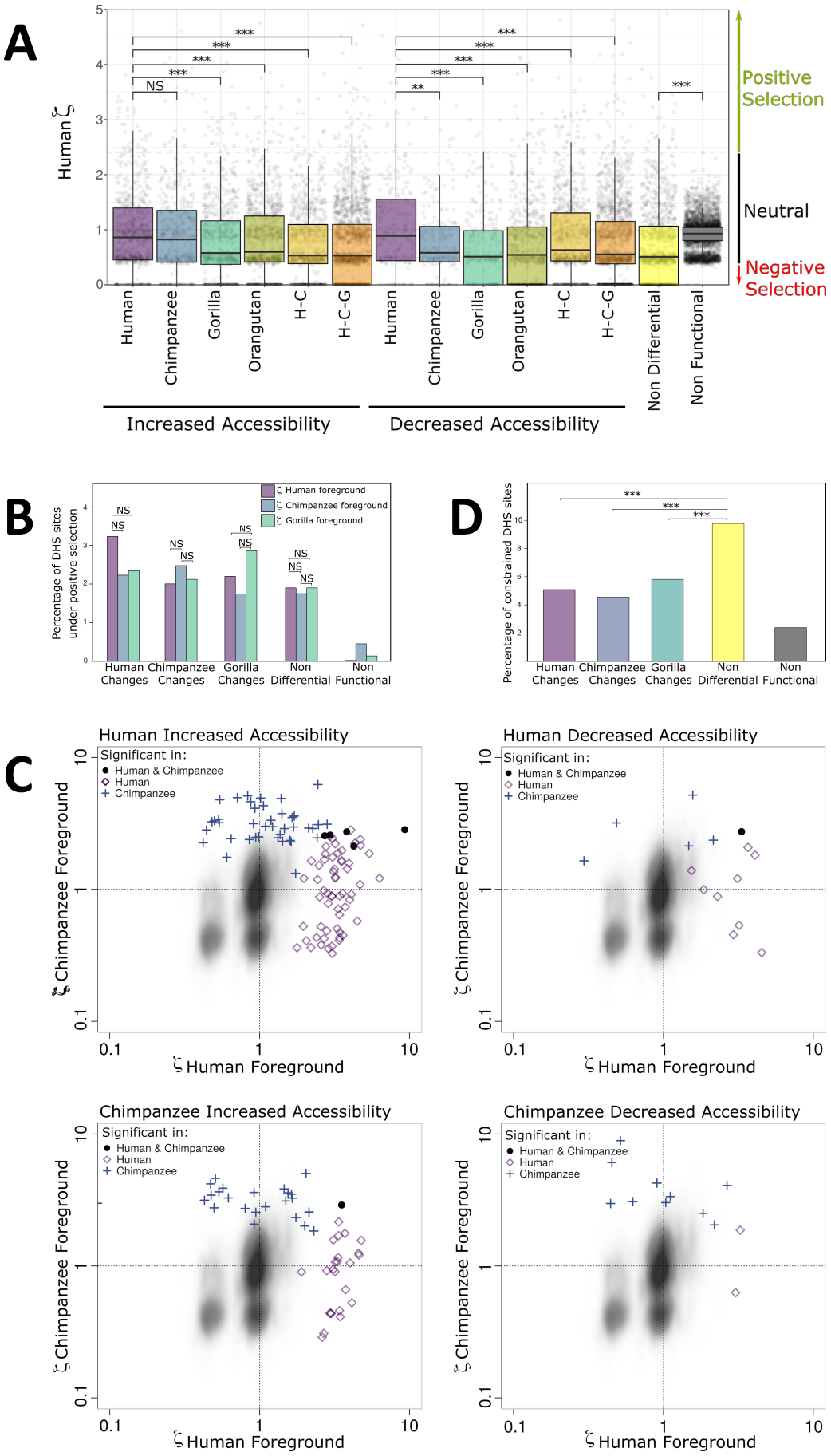

A
Human ζ
Positive Selection
Neutral
Negative Selection
Human
Chimpanzee
Gorilla
Orangutan
H-C
H-C-G
Non Differential
Non Functional
Increased Accessibility
Decreased Accessibility
B
Percentage of DHS sites under positive selection
ζ Human foreground
ζ Chimpanzee foreground
ζ Gorilla foreground
Human Changes
Chimpanzee Changes
Gorilla Changes
Non Differential
Non Functional
D
Percentage of constrained DHS sites
C
Human Increased Accessibility
Human Decreased Accessibility
Chimpanzee Increased Accessibility
Chimpanzee Decreased Accessibility
Significant in:
Human & Chimpanzee
Human
Chimpanzee
ζ Chimpanzee Foreground
ζ Human Foreground

TABLE 1

ALL DHS SITES		
	Total	**% of total**
Non differential (no changes)	53,078	59%
Differential (one or more changes)	36,666	41%
TOTAL	**89,744**	

SINGLE CHANGE AFFECTING SINGLE SPECIES				
	Increases	**Decreases**	**Total**	**% of differential**
Human	2,539	209	2,748	7%
Chimpanzee	1,539	196	1,735	5%
Gorilla	2,737	254	2,991	8%
Orangutan	3,759	646	4,405	12%
TOTAL	**10,574**	**1,305**	**11,879**	**32%**

SINGLE CHANGE AFFECTING MULTIPLE SPECIES				
	Increases	**Decreases**	**Total**	**% of differential**
Human/chimpanzee	1,735	814	2,549	7%
Human/chimpanzee/gorilla	1,736	1,453	3,189	9%
Human/chimpanzee/gorilla/ orangutan	1,317	4,992	6,309	17%
TOTAL	**4,788**	**7,259**	**12,047**	**33%**

MULTIPLE CHANGES AFFECTING TWO SPECIES				
	Increases	**Decreases**	**Total**	**% of differential**
Human/gorilla	997	221	1,218	3%
Human/orangutan	997	424	1,421	4%
Chimpanzee/gorilla	1,568	387	1,955	5%
Chimpanzee/orangutan	657	252	909	2%
Gorilla/orangutan	2,120	517	2,637	7%
TOTAL	**6,339**	**1,801**	**8,140**	**22%**

MULTIPLE CHANGES AFFECTING THREE SPECIES				
	Increases	**Decreases**	**Total**	**% of differential**
Human/chimpanzee/orangutan	794	906	1,700	5%
Human/gorilla/orangutan	691	554	1,245	3%
Chimpanzee/gorilla/orangutan	939	686	1,625	4%
TOTAL	**2,424**	**2,146**	**4,570**	**12%**

TABLE 2

HUMAN AND CHIMPANZEE CHANGES							
	Consistent (species-specific)		Consistent (multi-species)		Not consistent		Total
Human increases	245	72%	91	27%	6	2%	342
Chimpanzee increases	105	45%	114	49%	15	6%	234
Human decreases	21	14%	109	75%	16	11%	146
Chimpanzee decreases	17	17%	69	70%	12	12%	98

COMMON SITES							
Consistent (non differential)		Consistent (differential only in gorilla, orangutan)		Not consistent		Total
1,043	90%	57	5%	54	5%	1,154

FIG. S1

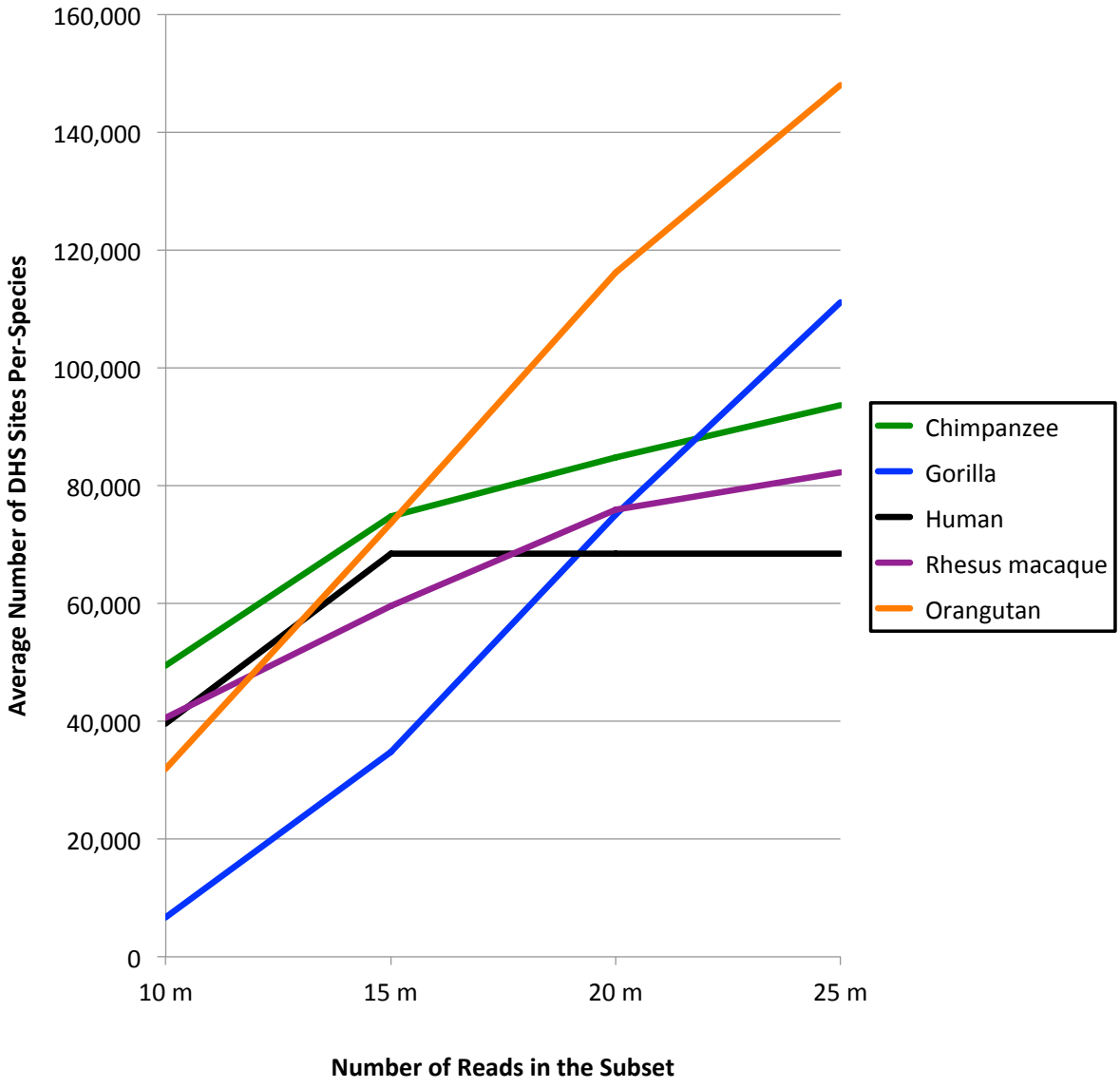

160,000
140,000
120,000
100,000
80,000
60,000
40,000
20,000
0
Average Number of DHS Sites Per-Species
10 m
15 m
20 m
25 m
Number of Reads in the Subset
Chimpanzee
Gorilla
Human
Rhesus macaque
Orangutan

FIG. S2

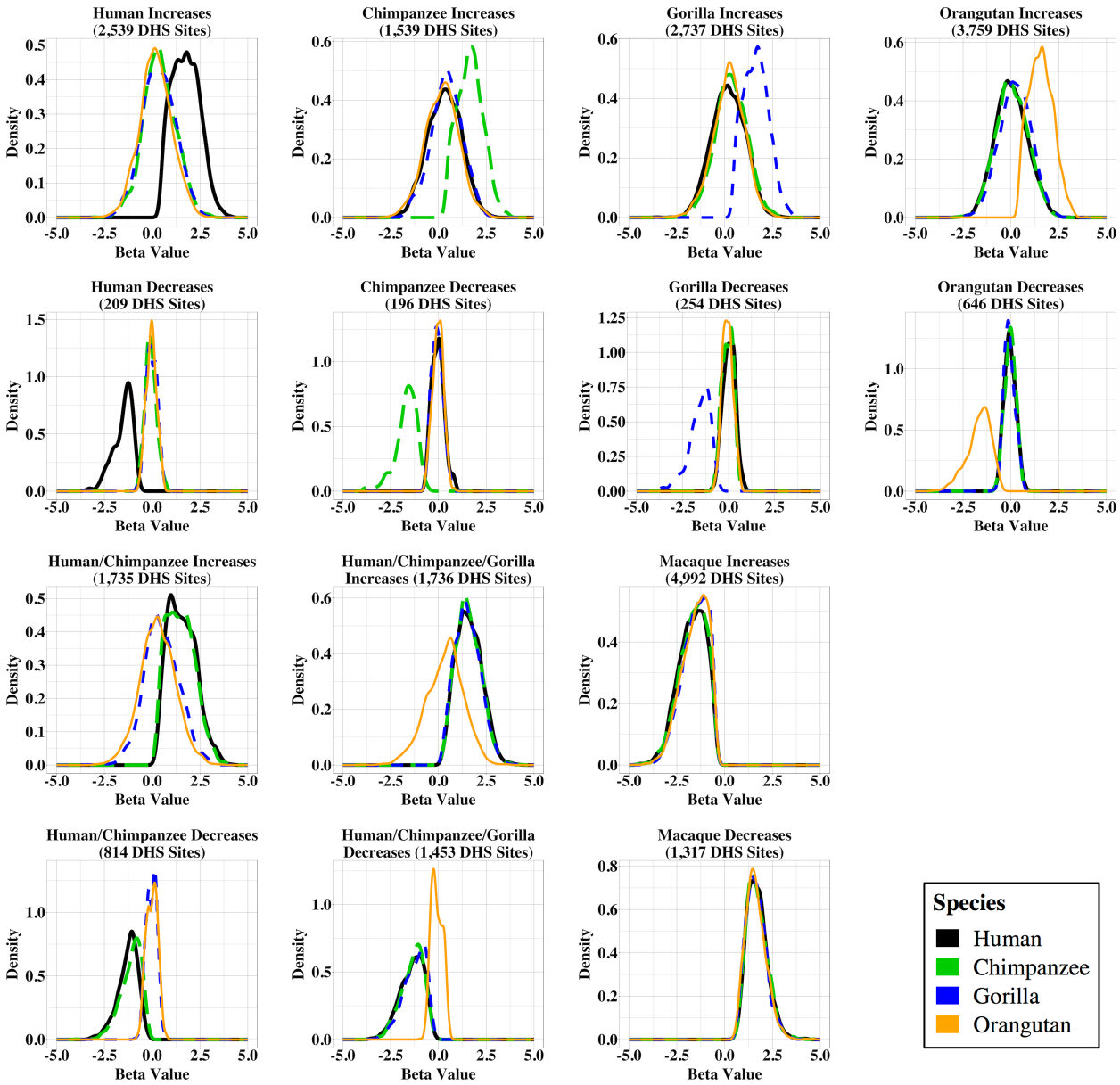

Human Increases (2,539 DHS Sites)
Chimpanzee Increases (1,539 DHS Sites)
Gorilla Increases (2,737 DHS Sites)
Orangutan Increases (3,759 DHS Sites)
Human Decreases (209 DHS Sites)
Chimpanzee Decreases (196 DHS Sites)
Gorilla Decreases (254 DHS Sites)
Orangutan Decreases (646 DHS Sites)
Human/Chimpanzee Increases (1,735 DHS Sites)
Human/Chimpanzee/Gorilla Increases (1,736 DHS Sites)
Macaque Increases (4,992 DHS Sites)
Human/Chimpanzee Decreases (814 DHS Sites)
Human/Chimpanzee/Gorilla Decreases (1,453 DHS Sites)
Macaque Decreases (1,317 DHS Sites)
Density
Beta Value
Species
Human
Chimpanzee
Gorilla
Orangutan

FIG. S3

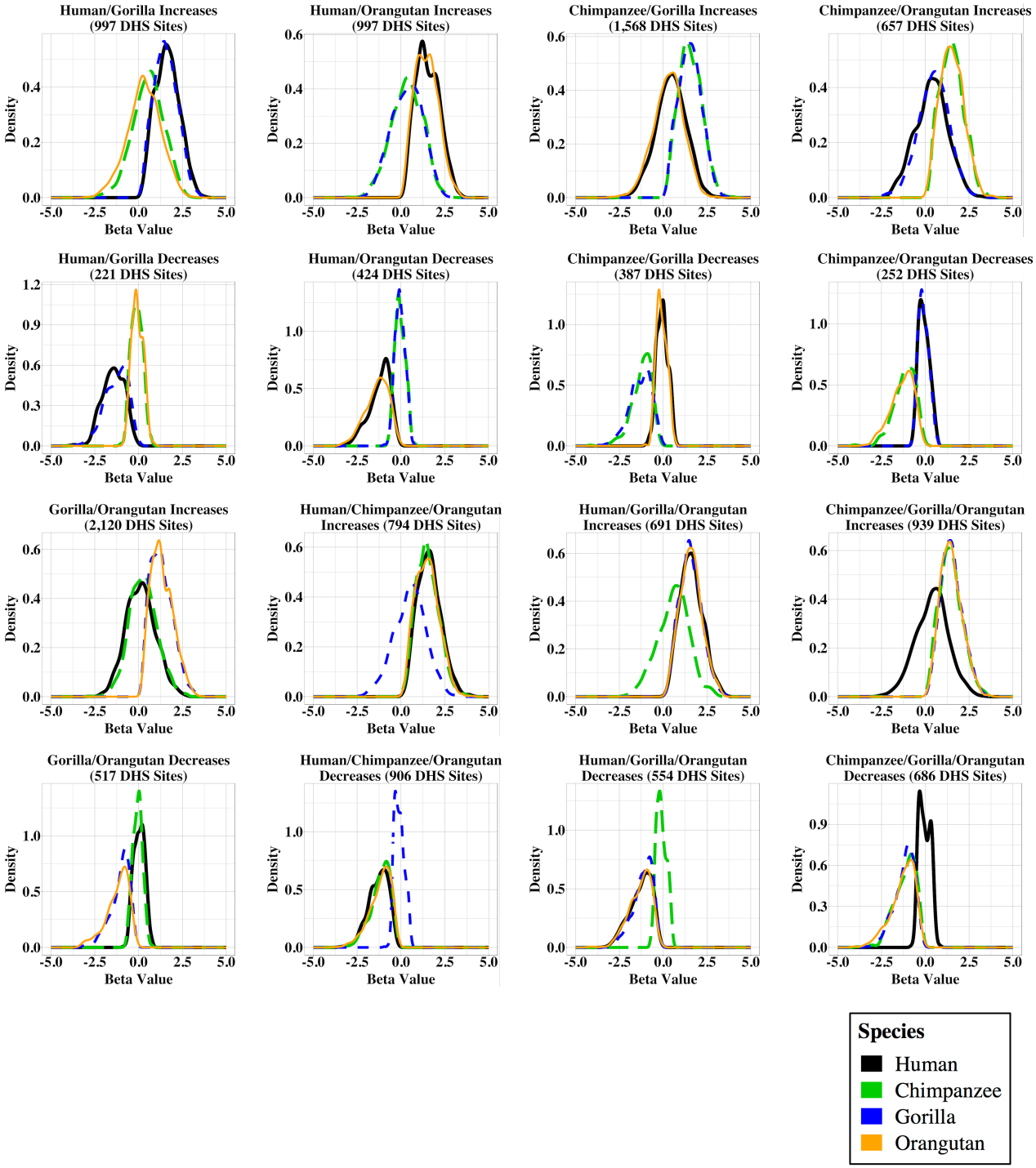

Human/Gorilla Increases (997 DHS Sites)
Human/Orangutan Increases (997 DHS Sites)
Chimpanzee/Gorilla Increases (1,568 DHS Sites)
Chimpanzee/Orangutan Increases (657 DHS Sites)
Human/Gorilla Decreases (221 DHS Sites)
Human/Orangutan Decreases (424 DHS Sites)
Chimpanzee/Gorilla Decreases (387 DHS Sites)
Chimpanzee/Orangutan Decreases (252 DHS Sites)
Gorilla/Orangutan Increases (2,120 DHS Sites)
Human/Chimpanzee/Orangutan Increases (794 DHS Sites)
Human/Gorilla/Orangutan Increases (691 DHS Sites)
Chimpanzee/Gorilla/Orangutan Increases (939 DHS Sites)
Gorilla/Orangutan Decreases (517 DHS Sites)
Human/Chimpanzee/Orangutan Decreases (906 DHS Sites)
Human/Gorilla/Orangutan Decreases (554 DHS Sites)
Chimpanzee/Gorilla/Orangutan Decreases (686 DHS Sites)
Density
Beta Value
Species
Human
Chimpanzee
Gorilla
Orangutan

FIG. S4

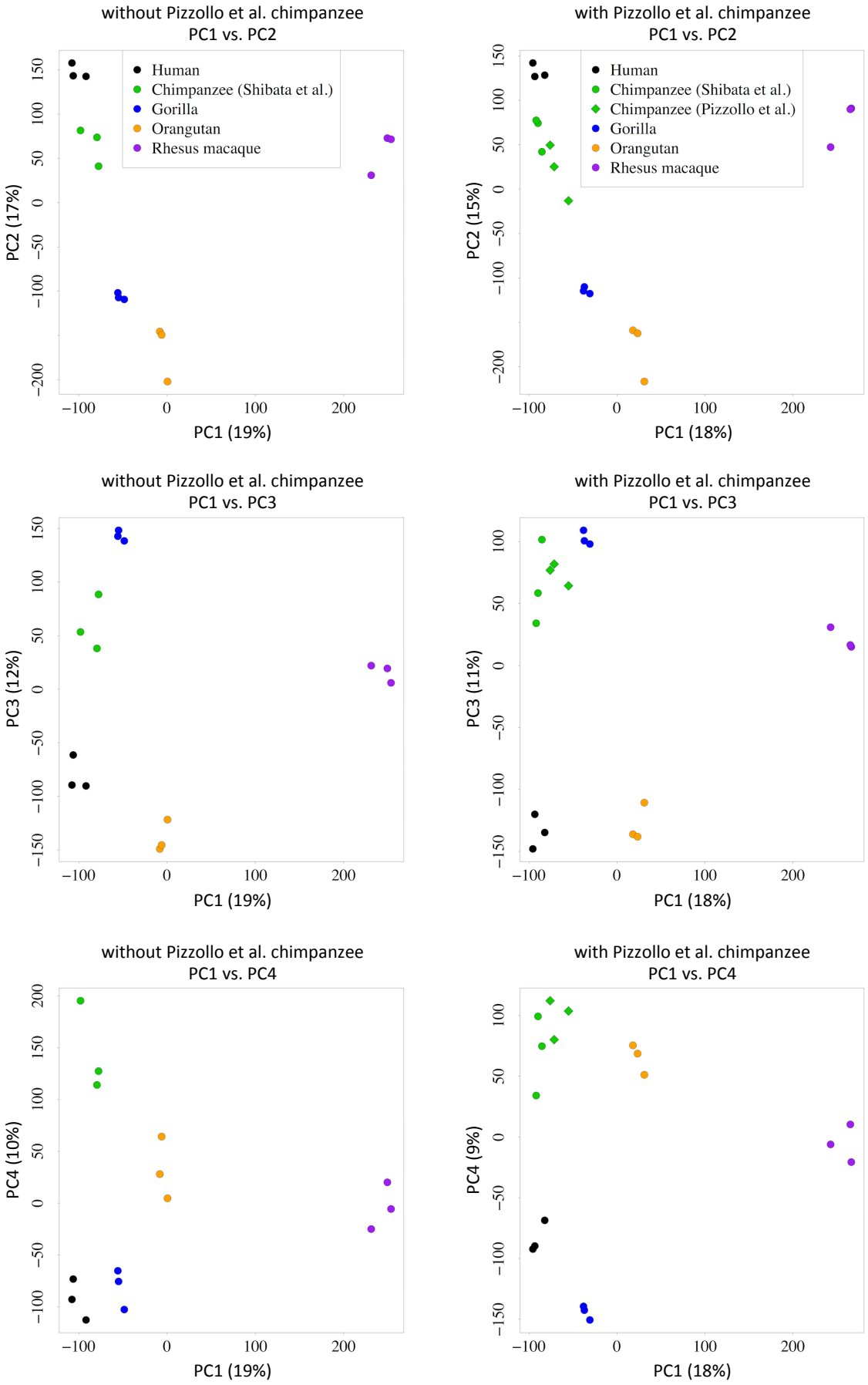

without Pizzollo et al. chimpanzee
PC1 vs. PC2
Human
Chimpanzee (Shibata et al.)
Gorilla
Orangutan
Rhesus macaque
PC2 (17%)
PC1 (19%)
with Pizzollo et al. chimpanzee
PC1 vs. PC2
Human
Chimpanzee (Shibata et al.)
Chimpanzee (Pizzollo et al.)
Gorilla
Orangutan
Rhesus macaque
PC2 (15%)
PC1 (18%)
without Pizzollo et al. chimpanzee
PC1 vs. PC3
PC3 (12%)
PC1 (19%)
with Pizzollo et al. chimpanzee
PC1 vs. PC3
PC3 (11%)
PC1 (18%)
without Pizzollo et al. chimpanzee
PC1 vs. PC4
PC4 (10%)
PC1 (19%)
with Pizzollo et al. chimpanzee
PC1 vs. PC4
PC4 (9%)
PC1 (18%)

TABLE S1

Sample	Coriell ID	Technical Replicates	Library Date	Number of Sequencing Lanes	Sequencer Type (Machine Name)	Biopsy Source	Sex	Age at Sampling	Notes
Chimpanzee 1	S003649	1	8/6/08	4 (2 lanes per sequencer)	Genome Analyzer II (HWI-EAS215) (HWUSI-EAS465)	Unknown	Male	9 YR	from the Yerkes Primates name is Josh
Chimpanzee 2	S003624	1	11/21/08	3 (2 lanes in one run; 1 lane in another)	Genome Analyzer IIx (HWUSI-EAS1676)	Unknown	Male	14 YR	from the Yerkes Primates name is Tank
Chimpanzee 3	S007603	1	11/21/08	4	Genome Analyzer II (HWUSI-EAS465)	Unknown	Male	16 YR	from the Yerkes Primates name is Donald
Chimpanzee 4	PR00238	1	9/13/12	1	HiSeq 2000 (D3NH4HQ1)	Unknown	Male	24 YR	name is Clint
Chimpanzee 5	S006007	1	9/13/12	1	HiSeq 2000 (D3NH4HQ1)	Unknown	Male	22 YR	
Chimpanzee 6	S008975	1	9/13/12	1	HiSeq 2000 (D3NH4HQ1)	Unknown	Male	19 YR	
Gorilla 1	PR00107	1	10/17/12	1	HiSeq 2000 (HWI-ST1293)	Unknown	Male	19 YR	
Gorilla 2	PR00669	1	10/17/12	1	HiSeq 2000 (HWI-ST1293)	Unknown	Male	15 YR	
Gorilla 3	PR01103	1	10/17/12	1	HiSeq 2000 (HWI-ST1293)	Unknown	Male	16 YR	
Human 1	AG16409	1	8/6/08	4 (2 lanes per sequencer)	Genome Analyzer II (HWI-EAS215) (HWUSI-EAS465)	Unknown	Male	12 YR	Caucasian
Human 2	GM02185	1	12/16/08	4	Genome Analyzer II (HWUSI-EAS465)	Unknown	Male	36 YR	Caucasian
Human 3	GM05879	1	12/16/08	4	Genome Analyzer II (HWUSI-EAS465)	Arm	Female	48 YR	Caucasian
Rhesus macaque 1	AG06252	2	8/6/08	6 (3 lanes per technical replicate)	Genome Analyzer IIx (HWUSI-EAS407)	Arm	Female	5 YR	Indian strain
Rhesus macaque 2	AG08305	2	12/16/08	7 total *Technical replicate #1* 4 lanes (3 lanes in one run, 1 lane in another) *Technical replicate #2* 3 lanes (2 lanes in one run, 1 lane in another)	*Technical replicate #1* Genome Analyzer II (HWUSI-EAS465) *Technical replicate #2* Genome Analyzer IIx (HWUSI-EAS1676) for 1 lane Genome Analyzer II (HWI-EAS215) for 2 lanes	Arm	Male	1 YR	
Rhesus macaque 3	AG08308	1	11/21/08	4 (3 lanes in one run; 1 lane in another)	Genome Analyzer II (unknown)	Arm	Male	1 YR	
Orangutan 1	PR00054	1	8/3/12	1	HiSeq 2000 (D3NH4HQ1)	Unknown	Male	4 YR	Sumatran orangutan (Pongo pygmaeus)
Orangutan 2	PR01107	1	8/3/12	1	HiSeq 2000 (D3NH4HQ1)	Unknown	Male	11 YR	Sumatran orangutan (Pongo pygmaeus)
Orangutan 3	GM04272	1	8/24/12	1	HiSeq 2000 (D3NH4HQ1)	Unknown	Male	10 YR	Sumatran orangutan (Pongo pygmaeus)

TABLE S2

		Tier 1 (No mismatches)						Tier 2 (1 mismatch)					
Replicate	Raw reads	Unique hits		Multiple hits		Unaligned		Unique hits		Multiple hits		Unaligned	
chimpanzee 1	50,096,931	31,674,269	63%	13,476,106	27%	4,946,556	10%	1,612,581	3%	1,735,149	3%	1,598,826	3%
chimpanzee 2	111,972,567	72,746,554	65%	29,841,886	27%	9,384,127	8%	2,750,866	2%	3,307,427	3%	3,325,834	3%
chimpanzee 3	34,388,687	21,848,911	64%	9,120,646	27%	3,419,130	10%	1,169,139	3%	1,379,792	4%	870,199	3%
gorilla 1	200,394,049	113,954,996	57%	46,452,406	23%	39,986,647	20%	17,164,167	9%	15,177,562	8%	7,644,918	4%
gorilla 2	207,690,245	116,195,086	56%	46,244,395	22%	45,250,764	22%	18,247,477	9%	15,755,405	8%	11,247,882	5%
gorilla 3	218,321,295	139,735,317	64%	56,153,448	26%	22,432,530	10%	7,057,645	3%	8,422,152	4%	6,952,733	3%
human 1	50,938,241	32,879,743	65%	12,283,795	24%	5,774,703	11%	2,136,098	4%	2,347,034	5%	1,291,571	3%
human 2	33,630,278	22,180,616	66%	8,140,672	24%	3,308,990	10%	1,141,379	3%	1,654,187	5%	513,424	2%
human 3	23,844,246	15,548,666	65%	5,556,547	23%	2,739,033	11%	1,017,236	4%	1,263,353	5%	458,444	2%
macaque 1	70,897,729	34,129,807	48%	12,911,828	18%	23,856,094	34%	9,092,082	13%	11,503,154	16%	3,260,858	5%
macaque 2	69,917,208	42,438,629	61%	14,314,693	20%	13,163,886	19%	4,609,783	7%	4,988,172	7%	3,565,931	5%
macaque 3	29,133,743	18,596,481	64%	6,046,944	21%	4,490,318	15%	1,456,525	5%	1,666,868	6%	1,366,925	5%
orangutan 1	204,329,161	126,205,138	62%	44,601,400	22%	33,522,623	16%	9,849,034	5%	10,301,994	5%	13,371,595	7%
orangutan 2	223,997,352	136,190,707	61%	51,142,667	23%	36,663,978	16%	10,465,410	5%	12,391,751	6%	13,806,817	6%
orangutan 3	228,118,747	124,383,888	55%	46,679,518	20%	57,055,341	25%	23,616,339	10%	20,069,852	9%	13,369,150	6%

	Final mapping results						Final results	
Replicate	Unique hits		Multiple hits		Unaligned		Unique monoclonal	
chimpanzee 1	33,286,850	66%	15,211,255	30%	1,598,826	3%	11,245,529	22%
chimpanzee 2	75,497,420	67%	33,149,313	30%	3,325,834	3%	51,812,496	46%
chimpanzee 3	23,018,050	67%	10,500,438	31%	870,199	3%	15,660,736	46%
gorilla 1	131,119,163	65%	61,629,968	31%	7,644,918	4%	108,696,016	54%
gorilla 2	134,442,563	65%	61,999,800	30%	11,247,882	5%	95,844,959	46%
gorilla 3	146,792,962	67%	64,575,600	30%	6,952,733	3%	119,415,564	55%
human 1	35,015,841	69%	14,630,829	29%	1,291,571	3%	14,476,479	28%
human 2	23,321,995	69%	9,794,859	29%	513,424	2%	15,287,873	45%
human 3	16,565,902	69%	6,819,900	29%	458,444	2%	13,516,977	57%
macaque 1	43,221,889	61%	24,414,982	34%	3,260,858	5%	40,813,669	58%
macaque 2	47,048,412	67%	19,302,865	28%	3,565,931	5%	23,197,615	33%
macaque 3	20,053,006	69%	7,713,812	26%	1,366,925	5%	13,836,905	47%
orangutan 1	136,054,172	67%	54,903,394	27%	13,371,595	7%	99,830,828	49%
orangutan 2	146,656,117	65%	63,534,418	28%	13,806,817	6%	47,027,229	21%
orangutan 3	148,000,227	65%	66,749,370	29%	13,369,150	6%	108,387,926	48%

Replicate	Mapped	After lift over (a)		Final counts (b)
chimpanzee 1	11,245,529	10,861,657	97%	10,624,634
chimpanzee 2	47,421,354	45,614,631	96%	44,529,183
chimpanzee 3	15,660,736	15,129,473	97%	14,783,843
gorilla 1	104,413,179	101,111,374	97%	98,813,074
gorilla 2	91,500,123	88,494,541	97%	86,567,493
gorilla 3	114,551,049	110,815,063	97%	108,405,287
human 1	14,388,372	14,388,372	100%	14,014,031
human 2	15,204,571	15,204,571	100%	14,810,459
human 3	13,437,729	13,437,729	100%	13,105,259
macaque 1	40,813,669	35,868,548	88%	34,430,825
macaque 2	23,197,615	20,611,242	89%	20,140,263
macaque 3	13,836,905	12,332,942	89%	12,040,473
orangutan 1	99,830,828	89,033,545	89%	86,998,499
orangutan 2	47,027,229	43,114,002	92%	42,076,972
orangutan 3	108,387,926	98,211,154	91%	95,803,771

TABLE S3

Sample	DHS sites
Chimpanzee 1	163,806
Chimpanzee 2	113,653
Chimpanzee 3	104,601
Gorilla 1	107,869
Gorilla 2	122,317
Gorilla 3	86,317
Human 1	99,030
Human 2	84,421
Human 3	90,871
Macaque 1	119,191
Macaque 2	123,237
Macaque 3	85,305
Orangutan 1	161,994
Orangutan 2	180,239
Orangutan 3	113,979

Species	DHS sites
Chimpanzee	75,700
Gorilla	63,512
Human	62,017
Macaque	68,720
Orangutan	94,883

Step	DHS Sites
Before filtering	166,331
After filtering for genomic coverage	124,900
After filtering for zero-count	89,744

TABLE S4

	Human Chimpanzee Gorilla Orangutan
human	[1 -¼ -¼ -¼]
chimpanzee	[-¼ 1 -¼ -¼]
gorilla	[-¼ -¼ 1 -¼]
orangutan	[-¼ -¼ -¼ 1]
macaque	[-¼ -¼ -¼ -¼]
human/chimpanzee	[½ ½ -⅓ -⅓]
human/gorilla	[½ -⅓ ½ -⅓]
human/orangutan	[½ -⅓ -⅓ ½]
chimpanzee/gorilla	[-⅓ ½ ½ -⅓]
chimpanzee/orangutan	[-⅓ ½ -⅓ ½]
gorilla/orangutan	[-⅓ -⅓ ½ ½]
human/chimpanzee/gorilla	[⅓ ⅓ ⅓ -½]
human/chimpanzee/orangutan	[⅓ ⅓ -½ ⅓]
human/gorilla/orangutan	[⅓ -½ ⅓ ⅓]
chimpanzee/gorilla/orangutan	[-½ ⅓ ⅓ ⅓]

TABLE S5

Step	Type of Change	Shibata et al. Pipeline	This Pipeline
Mapping	Mapping software	BWA	Bowtie
Mapping	Mapping parameters	Up to 3 mismatches	Up to 1 mismatch
Mapping	Mapping parameters	Multiple locations allowed	Unique locations only
Mapping	Chimpanzee genome	panTro2	panTro4
Mapping	Macaque genome	rheMac2	rheMac3
DHS site identification	Peak calling program	F-Seq	MACS2
DHS site identification	Per-species DHS sites	Union set of top 100,000 peaks present in any replicate	Union set of all peaks present in at least two replicates
DHS site identification	DHS site width	DHS sites less than 300 bases were extended to 300 bases	DHS site widths unchanged
DHS site identification	Windowing	DHS sites were divided into overlapping windows of 300 bases	None
DHS site filtering	Genomic coverage	None	Genomic coverage of at least 95% of the bases in every species
DHS site filtering	Zero-count filtering	At least one replicate had to have non-zero counts	At least two replicates had to have non-zero counts

	Original sites	With overlapping reads		With 1+ replicate peaks		With 2+ replicate peaks		Passed genomic coverage filter		Passed zero-count filter and analyzed		Not analyzed	
Human Increases	836	831	99%	761	91%	606	72%	531	64%	**310**	**37%**	526	63%
Chimpanzee Increases	676	650	96%	562	83%	442	65%	399	59%	**215**	**32%**	461	68%
Human Decreases	286	281	98%	N/A	N/A	N/A	N/A	198	69%	**133**	**47%**	153	53%
Chimpanzee Decreases	211	194	92%	N/A	N/A	N/A	N/A	135	64%	**89**	**42%**	122	58%

TABLE S6

INCREASED ACCESSIBILITY COMPARED TO RHESUS MACAQUE

	Human		Chimpanzee		Gorilla		Orangutan		H-C		H-C-G	
Proximal	62	2%	54	4%	235	9%	364	10%	44	3%	88	5%
Distal	2,107	83%	1,007	65%	1,638	60%	1,944	52%	1,433	83%	1,353	78%
Unannotated	370	15%	478	31%	864	32%	1,451	39%	258	15%	295	17%
Total	**2,539**		**1,539**		**2,737**		**3,759**		**1,735**		**1,736**	

DECREASED ACCESSIBILITY COMPARED TO RHESUS MACAQUE

	Human		Chimpanzee		Gorilla		Orangutan		H-C		H-C-G	
Proximal	38	18%	41	21%	27	11%	126	20%	274	34%	151	10%
Distal	103	49%	124	63%	183	72%	398	62%	299	37%	806	55%
Unannotated	68	33%	31	16%	44	17%	122	19%	241	30%	496	34%
Total	**209**		**196**		**254**		**646**		**814**		**1,453**	

	Non Differential	
Proximal	11,850	22%
Distal	30,371	57%
Unannotated	10,857	20%
Total	**53,078**	