Supplemental Information

Accelerated Evolution in Distinctive Species Reveals Candidate Elements for Clinically Relevant Traits, Including Mutation and Cancer Resistance

Elliott Ferris, Lisa M. Abegglen, Joshua D. Schiffman, and Christopher Gregg
Figure S1, Related to Figure 1. The majority of bat ARs are reproducible between closely related species and the distribution of autosomal elephant, hibernating bat, orca, dolphin, mole rat and squirrel ARs across human chromosomes is revealed. 

(A) Circle plot shows the genomic distribution of 50bp ARs for each of the six target species across the autosomes of the human genome. ARs are found on every human autosome for every species.

(B) Venn diagrams show the number of ARs shared between the closely related *Myotis lucifugus* (microbat) species and *Myotis davidii*, as well as the number shared between *Eptesicus fuscus* (big brown bat) and *Myotis lucifugus* (microbat). The data show that a
large number of ARs are shared between the closely related hibernating bat species, indicating reproducibility. We define the shared *Eptesicus fuscus* (big brown bat) and *Myotis lucifugus* (microbat) ARs as putative “hibernating bat ARs” (Hib Bat ARs).

(C) The plot shows the relationship between the number of ARs (log scale) for a given species versus the relative distance to its closest relative among the background species from which conserved elements are identified. Species with more ARs are more distant from their closest relative in the background based on the 4D neutral model ($r= +0.9$, $P=0.0041$, Pearson correlation).
Figure S2, Related to Figure 4. The genomic hotspot for elephant ARs at the FANCL locus is reproducible in the mm10-rooted multiple genome alignment.

(A) Phylogenetic tree of the lineages in our study of accelerated evolution in the elephant using the mm10-rooted multiple genome alignment. Foreground target species are indicated in colored text. Background species for the identification of conserved genomic elements are indicated in black.

(B) Venn diagram shows the number of elephant ARs identified in the hg19 rooted phylogenetic tree (Fig. 1A) versus the number in the mm10-rooted tree. The data show that 27% of ARs are shared across the different alignments and background species.

(C) FANCL is the gene associated with the largest number of hg19/mm10 reproducible elephant ARs, indicating that this relative hotspot is robust to different alignments and background species.
Figure S3, Related to Figure 5. BrainSpan atlas analysis of the developmental expression of coding and noncoding RNAs in *VRK2-FANCL-BCL11A* locus, as well as *SOX2* and *TP53*, in the human brain.

(A) Human BrainSpan atlas RNASeq datasets for the indicated genes revealed expression in at least some regions of the human brain at all ages. The heatmap is constructed from the mean expression level (FPKM, fragments per kilobase per millions reads) for each gene at each age across all brain regions. The expression level data are then centered by column (gene) and the heatmap colors reflect relatively high (red) versus low (blue) expression levels at different ages (pcw, post-conception weeks; mos, months after birth; yrs, years after birth). All genes show relatively higher expression during early stages of brain development.

(B) The correlation matrix shows the pairwise correlation of the RNASeq expression patterns of the genes in *VRK2-FANCL-BCL11A* locus, as well as *SOX2* and *TP53*, across the 524 brain regions and ages in the BrainSpan atlas. Only statistically significant positive correlations are shown (P<0.01, Pearson correlation). The data show that *FANCL* expression is significantly correlated to all genes in the matrix, consistent with the potential for regulatory relationships between these genes in the human brain.
Figure S4, Related to Figure 6. Elephant ARs at DNA damage response genes are putative functional regulatory elements that bind TFs and other regulatory proteins in diverse human cell types.

The heatmap shows the number of DNA damage response elephant ARs that are located in statistically significant human ChIP-Seq peaks (q-value <1x10^-5) for specific TFs and other regulatory proteins (y-axis) identified in each indicated cell class (x-axis). The CTCF enrichment data appear at the top and are highlighted in red, since CTCF binding sites are uniquely enriched in the elephant ARs in all of the major cell classes. The ARs overlap with binding sites for 311 different TFs or regulatory proteins across all cell classes and the largest enrichment effects are observed in blood cells. Note that not all TFs have been profiled in all human cell classes and therefore the absence of AR hits for a particular TF and cell type may be due to the absence of that TF in a particular study.
Figure S5, Related to Figure 7. Identification of high priority candidate functional elements for shaping squirrel pigmentation patterns at the MITF locus.

The genomic tracks show squirrel AR (green) located at the MITF locus, a major regulator of pigmentation and melanocyte development and proliferation. The neighboring lncRNA, SAMMSON, also has established roles in melanocyte biology. The blue track indicates 10 squirrel ARs for which the human homologs are DNAse I hypersensitive regions in human epidermal melanocytes. These are therefore priority candidate elements for regulatory changes that may contribute to the unique pigmentation patterns in the squirrel. The red track shows the subset of these elements that are binding sites for the transcription factor, TFAP2C, in human epidermal cells (all sites are significant peaks in the ChIP-Atlas, q < 1x10^-5). The data show how one can use our results to define candidate elements for a phenotype of interest (eg. pigmentation) and uncover candidate upstream regulatory pathways for functional studies.
SUPPLEMENTAL EXPERIMENTAL PROCEDURES

CONTACT FOR RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Dr. Christopher Gregg (chris.gregg@neuro.utah.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Whole blood samples from an adult female African elephant supplied by Utah’s Hogle Zoo (Salt Lake City, Utah) were used for RNA sequencing experiments. All experimental procedures were reviewed and approved by the Hogle Zoo’s ethical and scientific review board.

METHOD DETAILS

Identification of ARs in the Target Species

Our study is based on the species available in the hg19/GRCh37 vertebrate 100-way UCSC multiple genome alignment file (MAF). We extracted the data for the mammalian species in the MAF and then further refined the data, as follows. The target species (elephant, microbat, orca, dolphin, mole rat and squirrel) were selected to represent adaptations for diverse habitats (terrestrial, marine, aerial, subterranean) and lifestyles (hibernating and non-hibernating), and for specific biomedically relevant traits like cancer resistance, blood clot resistance and hibernation-related phenotypes. The background of species to define mammalian conserved regions of the genome were selected such that other hibernators, marine species, subterranean and aerial species were eliminated, so that the background species lacked the overt phenotypes of the target species. In addition, we pruned down the number of primate species included in the MAF to reduce the overall primate bias in the available UCSC dataset. The final background included a range of species, from marsupials to human, and represents all the major clades. From the background mammalian species, we identified 50 bp conserved regions using PhastCons (Hubisz et al., 2011; Siepel et al., 2005) with the following parameters: expected.length = 45, target.coverage = 0.3, rho= 0.31). To make the background conserved regions comparable across the six target species, we only included conserved regions that were also present in the 6 target species. Next, accelerated regions for each of the six target species were defined in the conserved regions using RPHAST (Hubisz et al., 2011; Pollard et al., 2010). Statistically significant ARs were defined with a false discovery rate threshold of 5%.

Analysis of GC-biased Gene Conversion Versus Adaptive Nucleotide Substitutions in Elephant ARs

We used previously published methods (Kostka et al., 2012) to test for biased gene conversion versus adaptive patterns of nucleotide substitutions at elephant ARs in the VRK2-FANCL-BCL11A hotspot. We analyzed the nucleotide substitution patterns in the elephant relative to the background species using the aggregated hg19-rooted multiple alignment file sequence for all 338 50 bp elephant ARs
in this hotspot. Using RPhast and a likelihood ratio test, we compared the log likelihood of two models examining: (1) lineage-specific GC-biased substitution rate compared to the background (gBGC model) versus (2) lineage-specific increased nucleotide substitutions without GC bias (adaptive model) compared to the background. We tested if the difference between log-likelihoods of the adaptive model – the GC-biased substitution model is positive, which indicates that adaptive selection is a better fitting model. Further, we test whether a model of adaptive selection + GC-biased conversion explains significantly more of the variance in nucleotide substitutions compared to GC-biased conversion alone. Our test statistic is two times the difference of the log likelihood of the two nested models with 1 degree of freedom and is expected to asymptotically follow a chi-squared distribution. We analyzed the ARs in aggregate rather than individually, because we expect that this approach results in more accurate modeling of substitution patterns across the hotspot ARs than an analysis of the individual 50 bp ARs.

Identification of ARs in Human and Mouse Putative Functional Genomic Elements
To test whether ARs are located in putative functional elements in the human and mouse genomes and characterize the biochemical activity of these elements, we used publically available DNase-Seq and ChIP-Seq datasets in the ChIP-Atlas (http://chip-atlas.org). Bedfiles of ARs were analyzed for human or mouse biochemically active genomic element enrichment effects in different tissues using in silico ChIP (http://chip-atlas.org/in_silico_chip) using the default significance threshold of 100. Enrichments were computed relative to 100x random permutations of the input data, according to the in silico ChIP algorithm documentation (https://github.com/inutano/chip-atlas/wiki#virtual_chip_doc). A significant enrichment effect was thresholded based on q < 0.05, providing an estimated 5% false discovery rate for enrichment effects. To identify ARs associated with specific classes of biochemically active elements, the bedfiles for particular experiments detailed in the text were downloaded from the ChIP-Atlas for the peak call files and are based on a significance threshold of q < 1x10^{-5} for a significant peak. Significant enrichment effects were frequently observed for multiple independent experiments annotated for the same cell class biochemical mark in the ChIP-Atlas. Our heatmaps display the data for the dataset with the maximum enrichment effect compared to random permutations of the data. The total number of datasets analyzed for DNase-Seq and ChIP-Seq for each cell class is shown in Table S2.

TFBS Motif Analysis in Elephant ARs in the VRK2-FANCL-BCL11A Region
We extracted the genome sequence for the elephant ARs across the VRK2-FANCL-BCL11A locus from the UCSC hg19 100-way vertebrate alignment file using Galaxy (Blankenberg et al., 2011). For each AR, the corresponding genomic sequence was extracted for the elephant, manatee, dog and human. To search for TFBS motifs in these sequences, we used the R package Biostrings (Pages et al., n.d.) to search for TFBS motifs (Jolma et al., 2013) curated as position weight matrices in the R package MotifDb (Shannon, 2014). A match was defined based on a threshold of 90% identity to the annotated PWM. For each of the 709 TFBS motifs annotated, we tallied the motif sites present in each AR in each of the four species. By contrasting these results, we were able to identify sites
present in all four species, sites that are present in the manatee, dog and human, but lost in the elephant. Additionally, we identified motifs present in the elephant, but absent in the other three species. Next, we scored the loss and gain of manatee-dog-human conserved motifs in the elephant for each TF in the ARs across the VRK2-FANCL-BCL11A region. For the lost motif score, we divide the number of conserved sites that are lost in the elephant by the mean number of sites that are present in the manatee, dog and human sequences in order to normalize the number of lost sites relative to the general prevalence of the motif. Similarly, we scored motifs appearing specifically in the elephant by dividing the number of elephant gained sites by the mean number of sites found in the other 3 species.

**Gene Ontology, Human Phenotype and Pathway Enrichment Analysis**

We tested the target species ARs for Human phenotype, biological process and MIG canonical pathway enrichments relative to the background of mammalian conserved elements using GREAT (hg19) (McLean et al., 2010). We used the basal plus 1 MB extension default settings. To link specific ARs to specific genes and human phenotypes, we used GREAT to assigned ARs to genes based hg19 annotations. AR assigned genes were then intersected with datasets downloaded the file ALL_SOURCES_ALL_FREQUENCIES_diseases_to_genes_to_phenotypes.txt from the Human Phenotype Ontology database (Köhler et al., 2017), which link gene mutations to specific human phenotypes, to create the final dataset.

We performed the gene ontology enrichment analysis for elephant DNA damage response genes in irradiated primary elephant blood cells and human pancreas T2D genes using ClueGO v2.3.2 (Bindea et al., 2009). Significant enrichments were thresholded at a Bonferroni step-down corrected P-value < 0.05 for a two-sided hypergeometric test. The Kappa score threshold was 0.4, GO fusion and GO grouping was set to True.

**Evaluation of AR Enrichments Within Elephant Differential Gene Expression Datasets**

To analyze AR enrichments at elephant DNA damage response genes we developed custom python code to assign ARs to genes according to the methods detailed for GREAT (McLean et al., 2010): Gene association rule: Basal+extension: 5000 bp upstream, 1000 bp downstream, 1000000 bp max extension. Elephant genes were mapped to hg19 human orthologs using the Bioconductor package Biomart (Durinck et al., 2009). Statistical enrichments for ARs at differentially expressed genes were calculated relative to the background of mammalian conserved regions in hg19 coordinates using the Bioconductor package LOLA (Sheffield and Bock, 2016).

**Transcriptome Profiling in Primary Elephant Peripheral Blood Lymphocytes**

Whole blood samples from a female African elephant supplied by Utah’s Hogle Zoo (Salt Lake City, Utah) were used for RNA sequencing experiments. Processing of blood for live cell experiments began within 1 hour of the blood being drawn. Peripheral blood
mononuclear cells (PBMCs) were isolated by Ficoll-Paque density-gradient followed by three washes in large volumes of PBS to remove platelets. Cells were cultured in RPMI 1640 (Life Technologies) supplemented with 20% filtered (0.22 µm filter) autologous plasma, penicillin/streptomycin, and L-glutamine. To enrich for lymphocytes, PBMCs were incubated at 37°C for 30 minutes in tissue culture treated dishes to allow monocytes to adhere. Suspension cells (peripheral blood lymphocytes) were harvested and counted. Cells were resuspended to 2x10⁶ cells/ml and either left untreated or exposed to 2 gray ionizing radiation (IR) in a RS-2000 X-ray Biological Research Irradiator. Cells were cultured (2ml/well) in 6 well plates at 37°C for 5 hours prior to RNA extraction. Cells were harvested, centrifuged to remove media, and cell pellets were frozen at -80°C. RNA was later extracted using PureLink RNA mini kit (Life Technologies) and Qiashredder (Qiagen) for homogenization. To remove genomic DNA, on column DNase treatment was performed using PureLink DNase (Life Technologies). RNA was concentrated using RNase MinElute Cleanup Kit (Qiagen). RNA concentrations were measured with Ribogreen (Life Technologies). RNA samples were submitted to the High Throughput Genomics Core at Huntsman Cancer Institute for RNA-sequencing where the Illumina TruSeq Stranded RNA kit with Ribo-Zero Gold was used to prepare the samples before HiSeq 125 cycle paired-end sequencing. The following adaptors were used (Read1: 5’-AGATCGGAAGAGCACACGTCTGAACTCCAGTCAG-3’ Read 2: 5’-AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT-3’).

The African elephant genome and annotations (loxAfr3, annotation build 78) were downloaded from Ensembl. P53 retrogenes that were not present in the Ensembl annotation were added manually. USeq_8.8.9beta MakeTranscriptome was used with max splices set to 10 Kb to identify the splice sites necessary to make a transcriptome index using novoindex. RNASEq reads were aligned to this transcriptome index using novoalign. DESeq2 (Anders et al., 2013) was used to calculate FPKM (fragments per kilobase per million reads) for annotated elephant genes (GenBank Accession GCA_000001905.1) and define differentially expressed genes. Differentially expressed genes are identified by comparing 1, 5, and 24 hours post radiation to time 0, or 1, 5, and 24 hours without radiation, respectively. Statistically significant differentially expressed genes were defined at a FDR threshold of 5%.

STATISTICAL ANALYSIS OF AR ENRICHMENTS AT DNA DAMAGE GENES
ARs and mammalian conserved regions (CRs) were assigned to genes based on custom code adopting the proximity algorithm approach of GREAT (McLean et al., 2010). The LOLA (Locus Overlap Analysis) package in R (http://databio.org/lola/) was then used to compute statistically significant AR enrichments at specific gene classes, such as 5hr and 24hr elephant DNA damage response genes, relative to the background of conserved regions. Significant enrichment effects were thresholded based on a p-value less than 0.05.

DATA AND SOFTWARE AVAILABILITY
A list of the software and datasets used and associated URLs or accession database IDs are provided here:
Software and Databases:

1. PHAST and RPHAST: [http://compgen.cshl.edu/phast/](http://compgen.cshl.edu/phast/); [http://compgen.cshl.edu/rphast/](http://compgen.cshl.edu/rphast/)
2. BioMart: [http://bioconductor.org/packages/release/bioc/html/biomaRt.html](http://bioconductor.org/packages/release/bioc/html/biomaRt.html)
3. LOLA: [https://bioconductor.org/packages/release/bioc/html/LOLA.html](https://bioconductor.org/packages/release/bioc/html/LOLA.html)
4. Biostrings: [http://bioconductor.org/packages/release/bioc/html/Biostrings.html](http://bioconductor.org/packages/release/bioc/html/Biostrings.html)
5. MotifDB: [http://bioconductor.org/packages/release/bioc/html/MotifDb.html](http://bioconductor.org/packages/release/bioc/html/MotifDb.html)
6. PWMEnrich: [http://bioconductor.org/packages/release/bioc/html/PWMEnrich.html](http://bioconductor.org/packages/release/bioc/html/PWMEnrich.html)
7. ClueGO: [http://apps.cytoscape.org/apps/cluego](http://apps.cytoscape.org/apps/cluego)
8. GREAT: [http://bejerano.stanford.edu/great/public/html/](http://bejerano.stanford.edu/great/public/html/)
9. GREAT-like: Gregg lab custom code for assigning ARs to genes based on the GREAT algorithm proximity approach
10. DESeq: [http://bioconductor.org/packages/release/bioc/html/DESeq.html](http://bioconductor.org/packages/release/bioc/html/DESeq.html)
11. NovoAlign: [http://www.novocraft.com/products/novoalign/](http://www.novocraft.com/products/novoalign/)
12. GALAXY: [https://usegalaxy.org](https://usegalaxy.org)
13. Circlize: [https://cran.r-project.org/web/packages/circlize/index.html](https://cran.r-project.org/web/packages/circlize/index.html)
14. SimpleSynteny: [http://www.SimpleSynteny.com](http://www.SimpleSynteny.com)
15. ChIP-Atlas: [http://chip-atlas.org](http://chip-atlas.org)
16. Custom python and R scripts: GREGG Lab - available upon request

Datasets:

1. The FASTQ and processed data files for the elephant primary blood cell RNASeq datasets have been deposited in GEO: GSE107117
2. UCSC hg19/GRCh37 100-way multiple alignment file: [http://hgdownload.cse.ucsc.edu/goldenPath/hg19/multiz100way/](http://hgdownload.cse.ucsc.edu/goldenPath/hg19/multiz100way/)
3. Human phenotype ontology genes-to-phenotype dataset (Build #127): [http://human-phenotype-ontology.github.io/downloads.html](http://human-phenotype-ontology.github.io/downloads.html)

REFERENCES

Anders, S., McCarthy, D.J., Chen, Y., Okoniewski, M., Smyth, G.K., Huber, W., Robinson, M.D., 2013. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat Protoc 8, 1765–1786. doi:10.1038/nprot.2013.099

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W.H., Pages, F., Trajanoski, Z., Galon, J., 2009. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics 25, 1091–1093. doi:10.1093/bioinformatics/btp101

Blankenberg, D., Taylor, J., Nekrutenko, A., Galaxy Team, 2011. Making whole genome multiple alignments usable for biologists. Bioinformatics 27, 2426–2428. doi:10.1093/bioinformatics/btr398

Durinck, S., Spellman, P.T., Birney, E., Huber, W., 2009. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomartR. Nat Protoc 4, 1184–1191. doi:10.1038/nprot.2009.97

Hubisz, M.J., Pollard, K.S., Siepel, A., 2011. PHAST and RPHAST: phylogenetic analysis with space/time models. Brief. Bioinformatics 12, 41–51. doi:10.1093/bib/bbq072

Jolma, A., Yan, J., Whittington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., Palin, K., Vaquerizas, J.M., Vincentelli, R., Luscombe, N.M., Hughes, T.R., Lemaire, P., Ukkonen, E., Kivioja, T., Taipale, J., 2013. DNA-Binding Specificities of Human Transcription Factors. Cell 152, 327–339. doi:10.1016/j.cell.2012.12.009

Kostka, D., Hubisz, M.J., Siepel, A., Pollard, K.S., 2012. The role of GC-biased gene conversion in shaping the fastest evolving regions of the human genome. Molecular Biology and Evolution 29, 1047–1057. doi:10.1093/molbev/msr279

Köhler, S., Vasilevsky, N.A., Engelstad, M., Foster, E., McMurry, J., Aymé, S., Baynam, G., Bello, S.M., Boerkoel, C.F., Boycott, K.M., Brudno, M., Buske, O.J., Chinnery, P.F., Cipriani, V., Connell, L.E., Dawkins, H.J.S., DeMare, L.E., Devereau, A.D., de Vries, B.B.A., Firth, H.V., Freson, K., Greene, D., Hamosh, A., Helbig, I., Hum, C., Jähn, J.A., James, R., Krause, R., F Laulederkind, S.J., Lochmüller, H., Lyon, G.J., Ogishima, S., Olry, A., Ouwehand, W.H., Pontikos, N., Rath, A., Schaefer, F., Scott, R.H., Segal, M., Sergouniotis, P.I., Sever, R., Smith, C.L., Straub, V., Thompson, R., Turner, C., Turro, E., Veltman, M.W.M., Vulliamy, T., Yu, J., Ziegenweidt, von, J., Zankl, A., Züchner, S., Zemojtel, T., Jacobsen, J.O.B., Groza, T., Smedley,
D., Mungall, C.J., Haendel, M., Robinson, P.N., 2017. The Human Phenotype Ontology in 2017. Nucleic Acids Res 45, D865–D876. doi:10.1093/nar/gkw1039

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., Bejerano, G., 2010. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol 28, 495–501. doi:10.1038/nbt.1630

Pages, H., Gentleman, R., Aboyoun, P., DeRoy, S., n.d. Biostrings: String objects representing biological sequences, and matching algorithms, 2008. R package version

Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., Siepel, A., 2010. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 20, 110–121. doi:10.1101/gr.097857.109

Shannon, P., 2014. MotifDb: An annotated collection of protein-DNA binding sequence motifs.

Sheffield, N.C., Bock, C., 2016. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. Bioinformatics 32, 587–589. doi:10.1093/bioinformatics/btv612

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., Weinstock, G.M., Wilson, R.K., Gibbs, R.A., Kent, W.J., Miller, W., Haussler, D., 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15, 1034–1050. doi:10.1101/gr.3715005