The UCSC Repeat Browser allows discovery and visualization of evolutionary conflict across repeat families

Jason D. Fernandes[1,2.3†], Maximilian Haeussler[1*#†], Armando Zamudio-Hurtado[1,4], W. James Kent[1], Sofie R. Salama[1,3#], David Haussler[1,2,3*#]

[1]Genomics Institute, University of California, Santa Cruz

[2]Department of Biomolecular Engineering, University of California, Santa Cruz

[3]Howard Hughes Medical Institute, University of California, Santa Cruz

[4]Big Data to Knowledge Program, California State University, Monterey Bay

[†]Co-first author

[#]Co-senior author

*correspondence to max@soe.ucsc.edu

ABSTRACT

Background

Nearly half the human genome consists of repeat elements, and many of these sequences play important roles in biological functions. However repeat elements pose several unique challenges to current bioinformatic analyses and visualization tools, as short repeat sequences can map to multiple genomic loci resulting in their misclassification and misinterpretation. Therefore, there is a continued need for standardized tools and techniques to interpret genomic data of repeats.

Results

We present the UCSC Repeat Browser, which consists of a complete set of repeat reference sequences derived from the gold standard repeat database RepeatMasker. The UCSC Repeat Browser contains mapped annotations from the human genome (hg19, due to the large existing annotation sets for this assembly) to these references, and presents all of them as a comprehensive interface to facilitate work with repetitive elements. Furthermore, it provides processed tracks of multiple publicly available datasets of biological interest to the repeat community, including ChIP-SEQ datasets for KRAB Zinc Finger Proteins (KZNFs) – a family of proteins known to bind and repress certain classes of repeats. Here we show how the UCSC Repeat Browser in combination with these datasets, as well as RepeatMasker annotations in several non-human primates, can be used to trace the independent trajectories of species-specific evolutionary conflicts.

Conclusions

The UCSC Repeat Browser allows easy and intuitive visualization of genomic data on consensus repeat elements, circumventing the problem of multi-mapping, in which sequencing reads of repeat elements map to multiple locations on the human genome. By developing a reference consensus, multiple datasets and annotation tracks can easily be overlaid to reveal complex evolutionary histories of repeats in a single interactive window. Specifically, we use this approach to retrace the history of several primate specific LINE-1 families across apes, and discover several species-specific routes of evolution that correlate with the binding and emergence of KZNFs.

Keywords

repeats; retroelements; genomics; krab zinc finger proteins; evolution;

INTRODUCTION

Transposable elements are significant drivers of eukaryotic genome evolution, and nearly half the human genome consists of these repetitive elements. Despite the high repeat content of the human genome, many genomic analyses struggle to deal with these regions as sequencing reads can often be assigned nearly equally well to multiple regions in the genome. Masking or filtering these reads is often considered a "conservative" approach in that it avoids mis-assigning the genomic location of a read, but it prevents the discovery of important biology occurring at repeat elements[1]. Indeed, many repeats already have established roles in important biological processes and complex behavioral phenotypes[2–4].

72

73 One of the major challenges in proper repeat-analysis is establishing a set of
74 standardized sequences, nomenclature and annotation sets that can be universally
75 understood by the scientific community. The most rigorous and commonly used
76 database of repeats are Repbase[5] and RepeatMasker[6]. Repbase began as a hand-
77 curated list in 1992 of 53 prototypic repeat sequences identified in the human genome[7].
78 By 2015, it contained more than 38,000 sequences in 134 species[5], making curation
79 and comprehension of each repeat family a daunting challenge. RepeatMasker is a
80 program that screens DNA (e.g. a newly sequenced genome) for repeat elements.
81 RepeatMasker classifies repeats by transforming the curated prototypic sequences in
82 Repbase database into consensuses that allow identification of a repeat family within a
83 genome. RepeatMasker also uses additional optimizations (e.g. building full length
84 repeat elements from smaller subparts, generalization (grouping together) of similar
85 elements, and specialization (using information about repeat structure)) designed to
86 improve the speed and quality of repeat detection (Figure 1A).

87

88 While the practice of repeat detection and classification is intimidating on its own, a
89 variety of tools exist to study repeats[8], the tools to dynamically visualize genomic data
90 and interact with existing annotation sets on repeats (e.g. protein coding regions,
91 conservation with other sequences and the list of matches in the genome) are currently
92 underdeveloped. Generating and mapping to a consensus version of individual repeats
93 has proven successful in illustrating novel biological features of retroelements, but has
94 largely been limited to static visualizations on targeted elements of interest[9,10].

Here we present the UCSC Repeat Browser. The Repeat Browser simplifies analysis of genomic data on repeats by providing automatically generated consensus sequences for all human repeat element classifications within RepeatMasker, and overlaying a precomputed set of comprehensive annotations in an interactive genomic browser environment (Figure 1A). Further, we demonstrate the utility of the Repeat Browser in uncovering and illustrating evolutionary conflict between a primate specific class of retrotransposons and their repressors.

IMPLEMENTATION

Generating Reference Sequences for Human Repeats

We first generated consensus reference sequences for each repeat family listed in the RepeatMasker annotation of the human genome (hg19). To do so, we downloaded all nucleotide sequences and their annotations in the RepeatMasker annotation track on the UCSC Human Genome Browser (hg19). We observed that extremely long repeats tended to represent recombination or misannotation events and therefore removed the longest 2% of sequences in all classes. We then aligned the 50 longest sequences of each class, as this produced a tractable number of sequences that allowed manual inspection of each alignment. Additionally, we decided to bias towards longer sequences for visualization purposes as this decreases the effect of insertions relative to the consensus which are otherwise invisible when plotted on shorter sequences. For each repeat family, these fifty sequences were realigned with MUSCLE[11] to create a consensus sequence. Each of these consensus sequences was then stored as a

118 “reference” in the Repeat Browser in a manner analogous to a single chromosome on

119 the Human Genome Browser. Each alignment is provided as a link in a “consensus

120 alignment” track for additional visual inspection by the user.

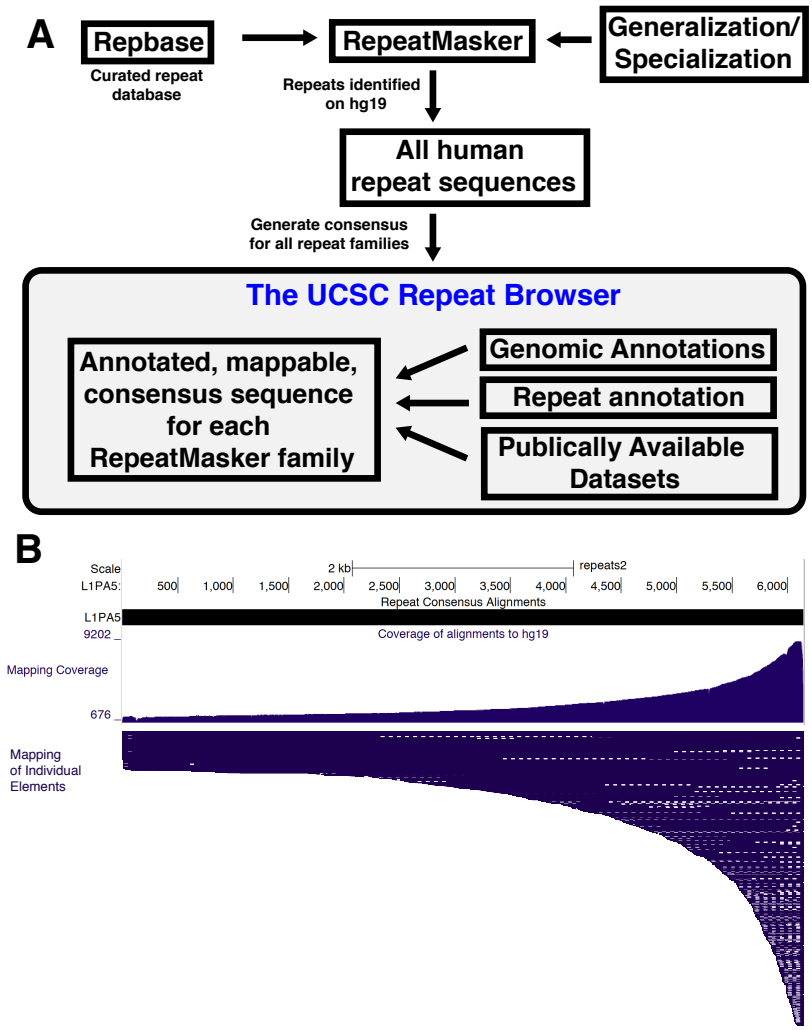

FIGURE 1: Building the UCSC Repeat Browser. A) Workflow for building the UCSC Repeat Browser. Repeat annotations and sequences are taken directly from RepeatMasker tracks across the human genome and used to build reference consensus sequences for every repeat family. Existing genomic annotations are then mapped to these consensuses. B) Every instance of a repeat family is mapped against the consensus; for certain families like L1PA5 individual instances will only contain a small portion of the consensus.

121

122 **Annotation of each repeat class**

123 For each repeat family, the consensus was mapped back to all of its repeats with

BLAT[12]. From this process, we generated a coverage plot illustrating the relative representation of the consensus from each genomic instance (Figure 1B). For example, the primate-specific L1PA5 element shows the expected distribution: most of the individual instances of the primate specific LINE-1 element, L1PA5, are short 3' truncations, meaning that most genomic loci annotated as L1PA5 do not contain the 5' portion; therefore the 3' end of the consensus is found relatively more often across the human genome (Figure 1B).We also ran Tandem Repeats Finder[13] and the EMBOSS ORF finder[14] on these consensus sequences in order to automatically annotate each consensus. We similarly aligned the RepeatMasker Peptide Library[15] with BLAST[16] and each of the original genomic sequences with BLAT[12] to each consensus.

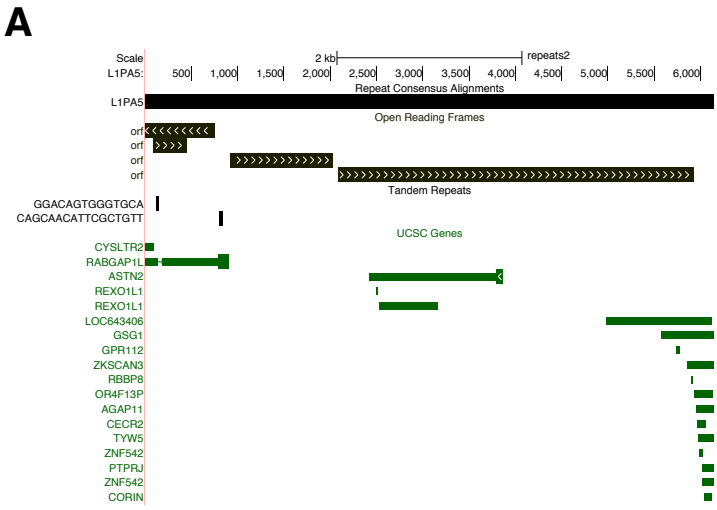

FIGURE 2: Mapping of existing annotations to the UCSC Repeat Browser. A) Existing hg19 annotation sets (e.g. UCSC Genes) that intersect RepeatMasker are lifted to the consensuses. Shown here are all genes that contain L1PA5 sequence as well as ORF structure.

Our alignment of individual repeat elements in the genome to their respective consensus sequence allows us to map any genome annotation to the genome consensus sequence, a process more generally known as "lifting". In this way, human

138 genes that contain repeat sequence (as annotated by GENCODE[17] and UCSC genes[18])
139 were “lifted” to each consensus sequences. The result of these procedures produces a
140 fully annotated and interactive consensus sequence that requires minimal prior
141 knowledge of the genomic organization of the repeat being analyzed and at the same
142 time allows lifting of any genome annotation (Fig 2).
143
144 **Mapping of Existing Genomic Datasets**
145 We also mapped genomic peaks bound by histone-modifying enzymes from ENCODE
146 datasets[19] as well as large-scale ChIP-SEQ collections KRAB Zinc Finger Proteins
147 (KZNFs) [20,21]. KZNFs are particularly compelling factors as they engage in evolutionary
148 “arms races” in which KZNFs evolve unique DNA binding properties to bind and repress
149 retrotransposons[9,22]. These retrotransposons then accumulate mutations that allow
150 evasion of KZNF-mediated repression[9]. In order to map this ChIP-SEQ data to the
151 Repeat Browser, we first downloaded raw fastqs from the Sequence Read Archive
152 (SRA)[23], mapped them to the reference genome (hg19) using bowtie2[24] and called
153 peaks using macs2[25] (Figure 3A). After this standard genomic mapping and peak
154 calling, we then took all peak summits that overlapped a repeat element as annotated in
155 the RepeatMasker track, extended them by 5 nt in both directions, and used BLAT to
156 map them to the appropriate (as determined by RepeatMasker annotation) RB
157 consensus sequence. In essence, this approach leverages each repeat instance as a
158 technical replicate with the mapping to the consensus representing an average of many
159 genomic “replicates’ (Figure 3A), as individual instances of a repeat family summits
160 produce a wide spread of mappings, but hundreds of them yield a clear average signal.

In order to determine the likely binding site within the consensus sequence we once again ran our peak caller (macs2) on the repeats to generate a list of “meta-summits” which represent the average location of each factor. We then generated a track which summarizes these meta-peaks for each consensus sequence allowing easy and quick determination of factors with correlated binding patterns (Figure 3B; visualized on Human Endogenous Retrovirus H (HERV-H)).

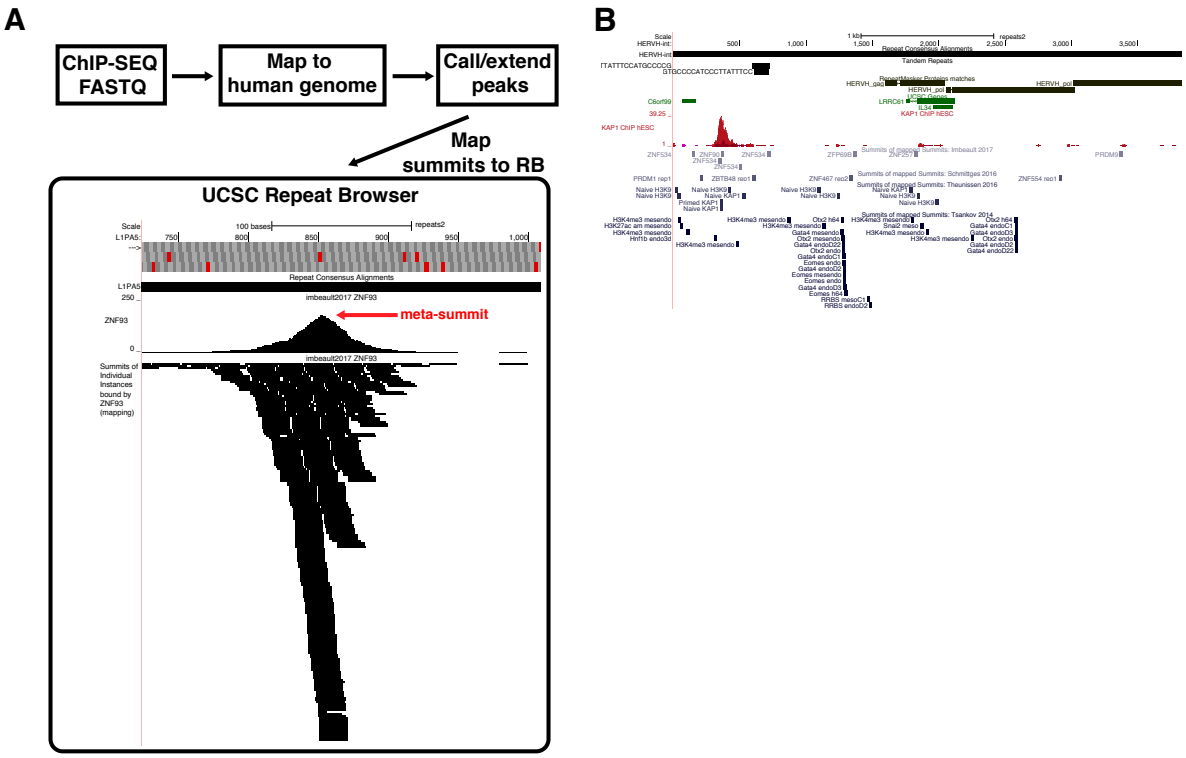

FIGURE 3: Mapping of KZNF ChIP-SEQ data to the UCSC Repeat Browser. A) Workflow for analyzing KZNF ChIP-SEQ. Data from existing collections was downloaded from SRA, analyzed via standard ChIP-SEQ workflows and the resulting summits mapped back to the RB for analysis. Mapping of individual summits produces a “meta-summit” (red arrow) that can be used for downstream analysis and which is stored separately in another annotation track. B) Example of an repeat family, HERVH-int (a primate endogenous retrovirus) with lifted annotations and datasets. Shown are tandem repeats, annotated ORFs gene tracks and ChIP-SEQ summits and meta-summits of a variety of transcription factors and histone marks.

RESULTS

170 **Comparative Analysis of L1PA elements**

171 In order to demonstrate the power of the UCSC Repeat Browser, we focused on recent
172 L1PA families. The L1PA lineage is a group of LINE-1 families specific to primates.
173 These elements are fully autonomous, encoding proteins (ORF1 and ORF2)
174 responsible for reverse transcription and re-integration of the retroelement (Fig 4A).
175 L1PA families evolve in bursts; higher numbers (e.g. L1PA17) indicate ancient
176 evolutionary origins, as evidenced by syntenic copies in distantly related species. Lower
177 numbers indicate more recent activity (e.g. L1PA1 is found only in humans (human-
178 specific) and is therefore also known as L1HS)[26]. We recently used the Repeat Browser
179 to demonstrate that multiple KZNFs will coordinate to battle L1PA elements and that
180 L1PA elements acquire multiple mutations to escape this concerted action (*Fernandes*
181 *et al, in preparation*). However, it is unclear if these mechanisms of escape are specific
182 to the human lineage, or occur independently in other primate lineages.

183

184 **Comparison of Primate Repeat Elements Reveals a Large Number of Gibbon**
185 **Specific L1PA4 Elements**

186 In order to trace the mechanisms of escape in other species, we downloaded the
187 complete sequences for every L1PA6 and younger repeat (as annotated in their UCSC
188 Genome Browser RepeatMasker tracks) in gibbon (nomLeu3), orangutan (ponAbe2),
189 chimp (panTro4), bonobo (panPan2) and human (hg19). We further restricted our
190 analysis to only full length elements by filtering out elements outside the expected size
191 range of an intact element. We then confirmed that RepeatMasker annotations matched

our expectations by comparing the expected evolutionary history of these families with the established species tree (Fig 4B).

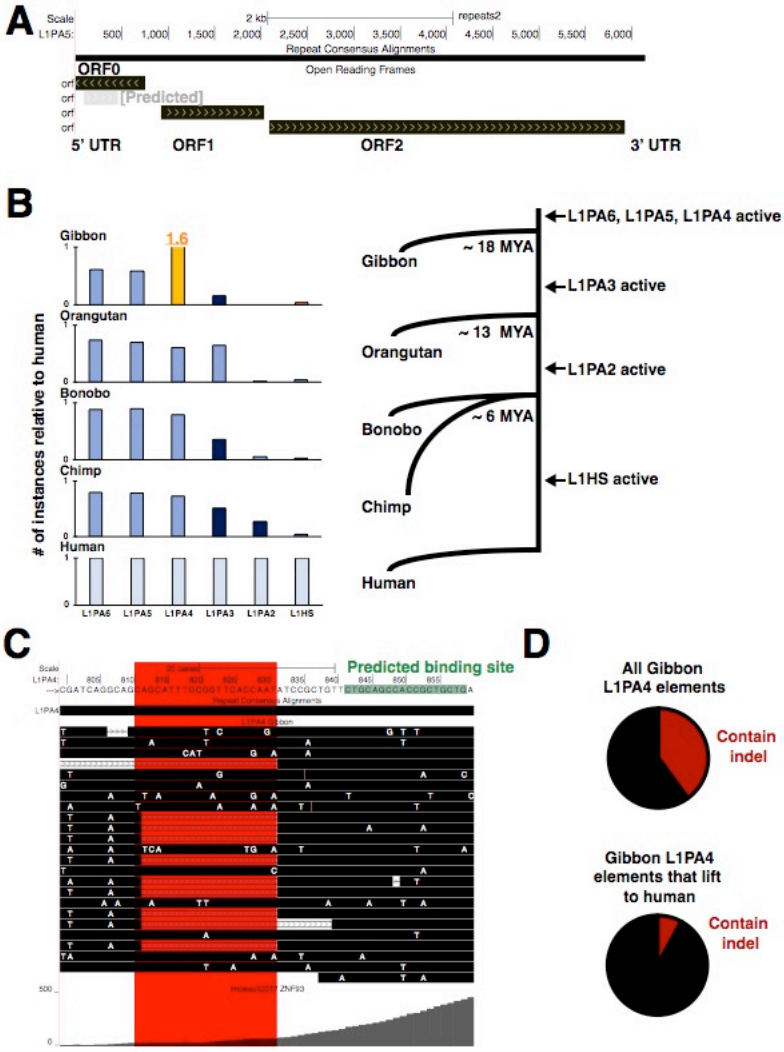

FIGURE 4: Comparative analysis of L1PA elements. A) L1PA5 elements consist of 3 ORFs (ORF1, ORF2 and ORF0 (antisense)) and 5' and 3' UTRs. B) Relative fraction of the number of instances of each L1PA family (6 to HS) detected by RepeatMasker in primate genomes compared to the human genome. Light blue coloring indicates similar frequencies to human, dark-blue indicates fewer than human, and gold (only in Gibbon L1PA4) indicates more than human. A cartoon (right) representing the evolutionary divergence of each species from the human lineage demonstrates that these fractions match the expected evolutionary history. C) Deletions in the Gibbon L1PA4 elements (red) are near the predicted ZNF93 binding site (green). D) Pie charts showing the fraction of all gibbon L1PA4 instances containing the gibbon-specific indels (top) or the instances shared between gibbon and human using liftOver to find syntenic L1PA4 instances.

As expected, the number of elements in older families were largely similar amongst all species that shared a common ancestor when the retrotransposon was active, while

196 human specific elements were found only (barring a few clear misannotations) in the
197 human genome (Fig 4B; white coloring on the left becomes bluer as species diverge).
198 The one exception is the L1PA4 family which has a marked increase in number of
199 instances in gibbon compared to human (Fig 4B; gold color). This is likely due to the
200 L1PA4 family expanding in a gibbon-specific fashion in a manner that still matches the
201 human L1PA4 classification. By using BLAT to lift each Gibbon L1PA4 instance to the
202 human Repeat Browser consensus (Fig 4C) we identify a subgroup of gibbon L1PA4
203 sequences that contain insertions and deletions near the ZNF93 binding site observed
204 in the great ape lineages. However, whereas young L1PA families (L1PA3, L1PA2,
205 L1HS) in great apes (orangutans, chimps, bonobos and humans) acquired a 129-bp
206 deletion to evade ZNF93, these gibbon-specific elements contain a much smaller
207 alterations to 20-bp region immediately 5' of the ZNF93 meta-summit, and near the
208 previous bioinformatically-predicted ZNF93 motif (Figure 4C), where the most C-
209 terminal fingers are predicted to contact. This suggests that ZNF93 in the human-
210 gibbon ancestor also repressed L1PA4 elements, and that gibbons acquired a separate
211 set of escape mutations than great apes (although it is possible that these changes are
212 driven by an unknown possibly gibbon-specific KZNF). Syntenic analysis of a subset of
213 Gibbon L1PA4 elements via liftOver supports the idea that these events are lineage-
214 specific (Figure 4D), as L1PA4 elements shared between gibbons and humans do not
215 contain these changes. Taken together, it appears that the large amount of gibbon
216 "L1PA4" actually consists of continued L1PA4 activity after divergence from the great
217 apes creating gibbon specific elements (e.g. L1 Gibbon-Specific or "L1GS" elements).
218 Thus, the gibbon and great ape lineages independently generated mutations that alter

219 the ZNF93 binding site, demonstrating the strength of the selective pressure KZNFs

220 exert on retroelements (Figure 4E).

221

222 **Novel Orangutan-Specific Deletions are Visible on the UCSC Repeat Browser**

223 We reasoned that L1PA4 elements provided strong signals of lineage-specific evolution

224 in gibbons since that particular L1PA family was active near the time which gibbons

225 diverged from humans.

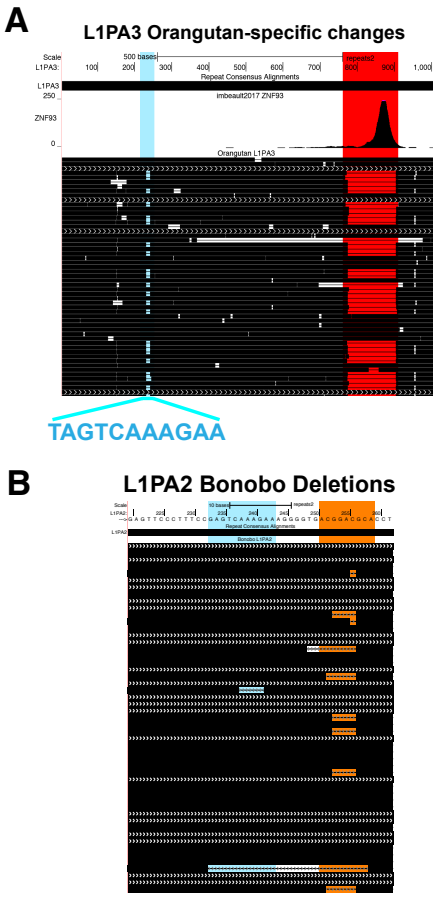

FIGURE 5: Lineage specific escape in the L1PA 5' UTR A) Orangutan L1PA3 elements acquire a 11 bp deletion (blue). This deletion occurs after the 129-bp ZNF93 binding site deletion (red). B) Bonobo L1PA2 elements mostly maintain the 11 bp region deleted in orangutans, but acquire deletions immediately adjacent to this region (orange).

226 Therefore, we decided to explore L1PA3 evolution in orangutans and L1PA2 evolution

227 in chimpanzees and bonobos. Lifting of orangutan L1PA3 elements on the Repeat

228 Browser L1PA3 consensus displayed a clear 11 nt deletion ~230 nt into the 5' UTR that
229 is not widespread in human, chimp or bonobo elements (Fig 5A). However, analysis of
230 existing KZNF ChIP-SEQ data, shows no specific factor that clearly correlates with this
231 deletion. It may be that we simply lack ChIP-SEQ of the appropriate factor (including the
232 possibility that the KZNF driving these changes evolved within the orangutan lineage) to
233 explain the evolutionary pattern seen in these orangutan-specific ("L1OS") elements.
234 Moreover, all instances containing this 11 nt deletion also harbor the 129 bp ZNF93
235 deletion, consistent with the model that this deletion occurred after orangutan L1PA3
236 elements escaped ZNF93. Interestingly, lifting of L1PA2 elements in bonobo, chimp
237 and human to the Repeat Browser show specific changes accumulating around ~250 nt
238 (human L1PA2 consensus numbering) with little evidence for indels near the orangutan-
239 specific deletion at 230 nt (Figure 5B). These observations are consistent with KZNF
240 targeting this overall region, but distinct species-specific escape mechanisms just as
241 observed in the ZNF93 case (Figure 6B).

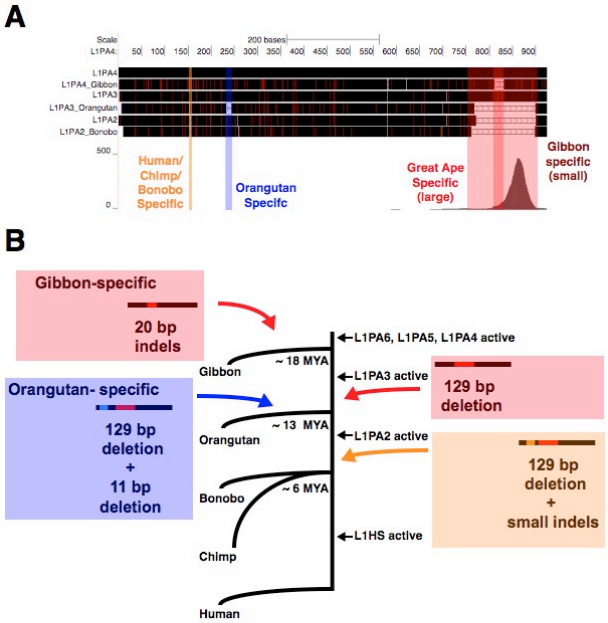

FIGURE 6: Model for L1PA evolution in different primate species. A) Representative sequences of L1PA families from human and other ape species lifted to the L1PA4 consensus show clear lineage-specific changes indicated by color. B) ZNF93 deletions (red) occur in 2 lineage-specific fashions: alterations to a 20 bp region in gibbon L1PA4 elements or a 129 bp deletion in L1PA3 elements of the great ape ancestor. Orangutans acquire a lineage-specific second deletion (11 bp, blue) in L1PA3 elements while the L1PA2 elements of bonobos, chimps and humans display a variety of changes immediately 3' of this deletion (orange).

DISCUSSION

The UCSC Repeat Browser provides an interactive and accessible environment to study repeat biology and side-steps the problem of mistakenly mapping reads to an incorrect locus by generating consensus representations of every repeat and focusing on how genome-wide datasets interact with repeat sequences independent of their genomic locus.

However, several caveats should be noted about Repeat Browser-based analyses. First, they rely entirely on RepeatMasker classifications (and in turn RepBase) and therefore depend on the quality of the annotations established in these collections. Second, the Repeat Browser uses its own consensus sequences to display genomic data, with these choices biased by length due to a preference for easy visualization over established biological relevance. However, custom installations of the browser allow the substitution of the pre-calculated sequence with a custom sequence. Indeed, we previously used this approach to create consensuses of L1PA3 subclasses (L1PA3long and L1PA3short (the great-ape specific 129bp deletion)) when tracing an evolutionary arms race between ZNF93 and L1PA3 elements.[9] Finally, the Repeat Browser and other consensus-based approaches risk diluting important, biologically

262 relevant signal driven by a few instances of a repeat type that may affect the cell by
263 virtue of their genomic location instead of their sequence. In these cases, the majority
264 of instances in these families may generate no signal and produce an underwhelming
265 "composite" Repeat Browser signal whereas an individual genomic locus may produce
266 a strong, reproducible, and functionally relevant signal. Therefore, we believe the
267 Repeat Browser analysis should be used in combination with existing genomic
268 approaches.
269
270 Finally, the existence of the UCSC Repeat Browser as a complete "genome" available
271 for download and manipulation should allow mapping and utilization of these set of
272 sequences with a variety of pre-existing tools. We expect that the repeat community will
273 make creative use of this tool beyond the workflows suggested here.
274

275 CONCLUSIONS

276 The UCSC Repeat Browser provides a fully interactive environment, analogous to the
277 Human Genome Browser, to study repeats. We show here that this environment
278 provides an intuitive visualization tool for analysis and hypothesis-generation, and, by
279 using the Repeat Browser, we demonstrate that sequence-specific changes in repeats
280 driven by KZNF binding can occur independently in different species. A test version of
281 the Repeat Browser is currently available at: http://bit.ly/repbrowser .
282
283
284 **Project name:** The UCSC Repeat Browser
285 **Project home page:** http://bit.ly/repbrowser

Operating system(s): Standard Web Browser

Programming language: Python, bash

License: Freely available for academic, nonprofit, and personal use.

Any restrictions to use by non-academics: No

FUNDING

This work was supported by EMBO ALTF 292-2011 and 4U41HG002371 to MH, and F32GM125388 to JDF. DH is an investigator of the Howard Hughes Medical Institution.

AUTHORS' CONTRIBUTIONS

MH developed the Repeat Browser with input from all other authors. JDF and AZ analyzed KZNF and repeat data. MH, JDF, SRS, WJK and DH conceived of the idea and contributed to the Repeat Browser's design.

ACKNOWLEDGEMENTS

We thank A. Smit and R. Hubley for helpful discussions about repeat consensus choice and annotations. We thank A. Ewing, J. Armstrong, F. Jacobs and D. Greenberg and all members of the Haussler lab for helpful comments and critical review of the manuscript.

REFERENCES

1. Slotkin, R. K. The case for not masking away repetitive DNA. *Mob. DNA* **9,** 15

309 (2018).
310 2. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable
311 elements: from conflicts to benefits. *Nat. Rev. Genet.* (2016).
312 doi:10.1038/nrg.2016.139
313 3. Pastuzyn, E. D. *et al.* The Neuronal Gene Arc Encodes a Repurposed
314 Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer. *Cell* **173,**
315 275 (2018).
316 4. Ding, Y., Berrocal, A., Morita, T., Longden, K. D. & Stern, D. L. Natural courtship
317 song variation caused by an intronic retroelement in an ion channel gene. *Nature*
318 **536,** 329–332 (2016).
319 5. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive
320 elements in eukaryotic genomes. *Mob. DNA* **6,** 11 (2015).
321 6. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0.
322 *http://www.repeatmasker.org*
323 7. Jurka, J., Walichiewicz, J. & Milosavljevic, A. Prototypic sequences for human
324 repetitive DNA. *J. Mol. Evol.* **35,** 286–291 (1992).
325 8. Goerner-Potvin, P. & Bourque, G. Computational tools to unmask transposable
326 elements. *Nat. Rev. Genet.* 1 (2018). doi:10.1038/s41576-018-0050-x
327 9. Jacobs, F. M. J. *et al.* An evolutionary arms race between KRAB zinc-finger
328 genes ZNF91/93 and SVA/L1 retrotransposons. *Nature* **516,** 242–5 (2014).
329 10. Sun, X. *et al.* Transcription factor profiling reveals molecular choreography and
330 key regulators of human retrotransposon expression. *Proc. Natl. Acad. Sci. U. S.*
331 *A.* **115,** E5526–E5535 (2018).

332 11. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high
333 throughput. *Nucleic Acids Res.* **32,** 1792–7 (2004).
334 12. Kent, W. J. BLAT---The BLAST-Like Alignment Tool. *Genome Res.* **12,** 656–664
335 (2002).
336 13. Benson, G. Tandem repeats finder: a program to analyze DNA sequences.
337 *Nucleic Acids Res.* **27,** 573–80 (1999).
338 14. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology
339 Open Software Suite. *Trends Genet.* **16,** 276–7 (2000).
340 15. Kohany, O., Gentles, A. J., Hankus, L. & Jurka, J. Annotation, submission and
341 screening of repetitive elements in Repbase: RepbaseSubmitter and Censor.
342 *BMC Bioinformatics* **7,** 474 (2006).
343 16. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local
344 alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
345 17. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The
346 ENCODE Project. *Genome Res.* **22,** 1760–74 (2012).
347 18. Hsu, F. *et al.* The UCSC Known Genes. *Bioinformatics* **22,** 1036–1046 (2006).
348 19. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the
349 human genome. *Nature* **489,** 57–74 (2012).
350 20. Imbeault, M., Helleboid, P.-Y. & Trono, D. KRAB zinc-finger proteins contribute to
351 the evolution of gene regulatory networks. *Nature* **543,** 550–554 (2017).
352 21. Najafabadi, H. S. *et al.* C2H2 zinc finger proteins greatly expand the human
353 regulatory lexicon. *Nat. Biotechnol.* **33,** 555–62 (2015).
354 22. Thomas, J. H. & Schneider, S. Coevolution of retroelements and tandem zinc

355 finger genes. *Genome Res.* **21,** 1800–12 (2011).

356 23. Leinonen, R., Sugawara, H., Shumway, M. & International Nucleotide Sequence
357 Database Collaboration. The sequence read archive. *Nucleic Acids Res.* **39,** D19-
358 21 (2011).

359 24. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat.*
360 *Methods* **9,** 357–359 (2012).

361 25. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9,**
362 R137 (2008).

363 26. Khan, H., Smit, A. & Boissinot, S. Molecular evolution and tempo of amplification
364 of human LINE-1 retrotransposons since the origin of primates. *Genome Res.* **16,**
365 78–87 (2006).

366