GigaScience

RepeatFiller newly identifies megabases of aligning repetitive sequences and improves annotations of conserved non-exonic elements

--Manuscript Draft--

Manuscript Number:	GIGA-D-19-00249R1
Full Title:	RepeatFiller newly identifies megabases of aligning repetitive sequences and improves annotations of conserved non-exonic elements
Article Type:	Technical Note
Funding Information:	Max-Planck-Gesellschaft (-) — Dr. Michael Hiller Leibniz-Gemeinschaft (SAW-2016-SGN-2) — Dr. Michael Hiller
Abstract:	Transposons and other repetitive sequences make up a large part of complex genomes. Repetitive sequences can be co-opted into a variety of functions and thus provide a source for evolutionary novelty. However, comprehensively detecting ancestral repeats that align between species is difficult since considering all repeat-overlapping seeds in alignment methods that rely on the seed-and-extend heuristic results in prohibitively high runtimes. Here, we show that ignoring repeat-overlapping alignment seeds when aligning entire genomes misses numerous alignments between repetitive elements. We present a tool – RepeatFiller – that improves genome alignments by incorporating previously-undetected local alignments between repetitive sequences. By applying RepeatFiller to genome alignments between human and 20 other representative mammals, we uncover between 22 and 84 megabases of previously-undetected alignments that mostly overlap transposable elements. We further show that the increased alignment coverage improves the annotation of conserved non-exonic elements, both by discovering numerous novel transposon-derived elements that evolve under constraint and by removing thousands of elements that are not under constraint in placental mammals. In conclusion, RepeatFiller contributes to comprehensively aligning repetitive genomic regions, which facilitates studying transposon co-option and genome evolution.
Corresponding Author:	Michael Hiller GERMANY
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	
Corresponding Author's Secondary Institution:	
First Author:	Ekaterina Osipova
First Author Secondary Information:	
Order of Authors:	Ekaterina Osipova Nikolai Hecker Michael Hiller
Order of Authors Secondary Information:	
Response to Reviewers:	We have uploaded the point-by-point response as a separate pdf document in the category supplementary Data, but labeled as 'PointByPointResponse'.
Additional Information:	
Question	**Response**
Are you submitting this manuscript to a special series or article collection?	No

Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	Yes
Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	Yes
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript. Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	Yes

RepeatFiller newly identifies megabases of aligning repetitive sequences and improves annotations of conserved non-exonic elements

Ekaterina Osipova [1,2,3], Nikolai Hecker [1,2,3], Michael Hiller [1,2,3*]

[1]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany
[2]Max Planck Institute for the Physics of Complex Systems, Dresden, Germany
[3]Center for Systems Biology Dresden, Germany

ORCID IDs:
Nikolai Hecker: 0000-0003-1693-4257; Michael Hiller: 0000-0003-3024-1449

*To whom correspondence should be addressed:
Michael Hiller
Computational Biology and Evolutionary Genomics, Max Planck Institute of Molecular Cell Biology and Genetics & Max Planck Institute for the Physics of Complex Systems, Dresden, Germany.
Tel: +49 351 210 2781
Fax: +49 351 210 1209
Email: hiller@mpi-cbg.de

Running title: RepeatFiller newly identifies aligning repetitive elements

Keywords: transposons, conserved non-exonic elements, genome alignments

Abstract

Transposons and other repetitive sequences make up a large part of complex genomes. Repetitive sequences can be co-opted into a variety of functions and thus provide a source for evolutionary novelty. However, comprehensively detecting ancestral repeats that align between species is difficult since considering all repeat-overlapping seeds in alignment methods that rely on the seed-and-extend heuristic results in prohibitively high runtimes. Here, we show that ignoring repeat-overlapping alignment seeds when aligning entire genomes misses numerous alignments between repetitive elements. We present a tool – RepeatFiller – that improves genome alignments by incorporating previously-undetected local alignments between repetitive sequences. By applying RepeatFiller to genome alignments between human and 20 other representative mammals, we uncover between 22 and 84 megabases of previously-undetected alignments that mostly overlap transposable elements. We further show that the increased alignment coverage improves the annotation of conserved non-exonic elements, both by discovering numerous novel transposon-derived elements that evolve under constraint and by removing thousands of elements that are not under constraint in placental mammals. In conclusion, RepeatFiller contributes to comprehensively aligning repetitive genomic regions, which facilitates studying transposon co-option and genome evolution.

Source code: https://github.com/hillerlab/GenomeAlignmentTools

Introduction

A substantial portion of vertebrate genomes consist of transposons and other repetitive sequences [1, 2]. While most repeats are estimated to evolve neutrally [3], transposons are important substrates for evolutionary tinkering [4, 5]. For example, transposon-derived sequences contribute to the transcriptome by providing alternatively spliced exons [6, 7]. By contributing transcription factor binding sites, promoters, and distal regulatory elements, co-opted transposons are involved in rewiring of regulatory networks and drive regulatory innovation [7-15]. Importantly, a sizeable portion of evolutionarily constrained regions arose from ancestral transposon sequences [16, 17]. Studying how ancestral transposons and other repeats were co-opted into functional roles requires whole genome alignments that comprehensively align orthologous repeats.

The nature of repetitive sequences such as transposons, however, leads to many paralogous alignments, which pose a challenge for comprehensively aligning orthologous repeats between vertebrate genomes. Most methods for aligning entire genomes use a seed-and-extend heuristic, originally implemented in BLAST [18], to find local alignments between the sequences of two genomes. The seeding step of this heuristic detects short words or patterns (called seeds) that match between the sequences of the two genomes. This can be computed very efficiently. Seed detection is then followed by a computationally more expensive alignment extension step that considers ungapped and gapped local alignments. Given that repetitive sequences provide numerous seed matches to paralogous repeat copies in a whole genome comparison, it is computationally infeasible to start a local alignment from seeds located in repetitive sequences. Therefore, seeds that overlap repetitive regions are not used to start a local alignment phase, either by masking repetitive regions before aligning genomes [19-22] or by dynamically adapting seeding parameters by the observed seed frequencies [23]. Consequently, alignments between repeats are only found during the extension phase, initiated from seeds outside the repeat boundaries. This can be problematic if the regions flanking a repeat have been diverged to an extent that no seed in the vicinity of the repeat can be found.

Here, we investigated to which extent aligning repetitive sequences are missed in whole genome alignments. We show that ignoring repeat-overlapping seeds misses between 22 and 84 Mb of mostly repetitive elements that actually align between mammals and we provide a tool, called RepeatFiller, to incorporate such repeat-overlapping alignments into genome alignments. We further show that a subset of aligning sequences detected by RepeatFiller evolve under evolutionary constraint, which uncovers previously-unknown conserved non-exonic elements and thus improves the annotation of constrained elements.

Results

RepeatFiller incorporates several megabases of aligning repetitive sequences to mammalian genome alignments

To investigate how many aligning repetitive elements have been missed in alignments between mammalian genomes, we adopted a previously-developed approach that was initially devised to detect novel local alignments between a pair of distantly-related species [24, 25]. The original approach focused on unaligning regions that are flanked by aligning blocks in co-linear alignment chains [26], which are detected in the first all-vs-all genome alignment step. In a second step, this original approach used lastz [21] with highly-sensitive seeding and (un)gapped extension parameters to align the previously-unaligning regions again. This second round of highly-sensitive local alignment can uncover novel alignments that are co-linear with already-detected alignment blocks. Here, we adopted this approach by introducing two key changes. First, we increased alignment parameter sensitivity only slightly, but unmasked the unaligning region. This implies that all seeds, including repeat-overlapping seeds, will be considered (Figure 1). By restricting the size of the unaligning regions to smaller regions of at most 20 kb, we reason that novel local alignments detected with a similar sensitivity level likely constitute orthologous alignments. Second, while the previous approach computed all alignment chains again from scratch using previously-detected and novel local alignments, our new approach directly adds novel alignments to existing alignment chains, thus removing the need for a chain re-computing step. This approach is called RepeatFiller and is available at https://github.com/hillerlab/GenomeAlignmentTools.

To investigate how many aligning repetitive elements can be added by RepeatFiller, we built alignment chains between the human (hg38) genome assembly and the genomes of 20 other mammals that represent the major mammalian clades (Figure 2A, Supplementary Table 1). We found that RepeatFiller adds between 22.4 Mb (rhesus macaque) and 83.7 Mb (rabbit) of aligning sequence, which represents between 0.7 – 2.6% of the human genome (Figure 2A, Supplementary Table 1). RepeatFiller added fewer new alignments for the rhesus macaque likely because the genomes of both species are very similar (their evolutionary distance is less than 0.07 substitutions per neutral site). This makes it more likely to find seeds outside of masked repetitive regions and to extend alignments into repeats during the extension phase. By overlapping the new alignments with repetitive elements annotated in the human genome, we found that the vast majority of newly-aligned sequences overlap repeats, in particular transposable elements (Figure 2A, Supplementary Table 1). The runtime of the RepeatFiller step is between 14.7 and 43.4 CPU hours (Supplementary Table 1), and thus adds little to the

runtime of the initial genome-wide all-vs-all pairwise alignment step that is typically around ~1000 CPU hours.

Next, we investigated what factors are associated with differences in the amount of newly-aligned sequences per species. In these tests, we excluded rhesus macaque that is closely related to human as an outlier. First, as expected, the percent of the genome that is repeat-masked significantly influences the number of newly-aligned bases (P=0.0004, Supplementary Table 1), which supports our assumption that the initial alignment step misses alignments due to repeat-masking rather than sequence dissimilarity. Second, we investigated how assembly contiguity of the query genome influences the results. We found that the scaffold N50 value has a small but non-significant effect on the amount of added aligned bases (P=0.067, Supplementary Table 1). Since chains cannot span scaffold boundaries, we further tested the influence of scaffold N50 values by applying RepeatFiller to alignments of three fragmented mammalian assemblies: Parnell's mustached bat, rock hyrax and kangaroo rat, which have scaffold N50 values between 23 and 36 kb. While RepeatFiller still added a substantial amount of new alignments, ranging from 32 Mb to 35 Mb (Figure 2B, Supplementary Table 2), more new alignments were generally found for more contiguous mammalian assemblies. Together, this shows that a considerable portion of aligning transposon sequences are missed when repeat-overlapping seeds are ignored and that for both fragmented or contiguous mammalian genomes RepeatFiller can detect such alignments with little extra computational runtime.

RepeatFiller also detects additional alignments for non-mammalian genomes

The majority of the newly-detected alignments between mammalian genomes overlap transposable elements or other repeats. One would therefore expect that RepeatFiller application to alignments of species with less repeat-rich genomes detects fewer novel alignments. To test whether this is generally true, we applied RepeatFiller to alignments of birds (zebra finch aligned to chicken), reptiles (green anole aligned to bearded dragon; American alligator aligned to painted turtle), and insects (*Drosophila pseudoobscura* aligned to *D. melanogaster*). For birds and insects, whose genomes generally consist of <20% repeats [27-29], RepeatFiller added few new alignments (1.9 Mb for birds representing 0.18% of the chicken genome, 231 kb for Drosophilids representing 0.16% of the *D. melanogaster* genome) (Figure 3, Supplementary Table 2). For reptiles, RepeatFiller added 4.5 Mb of new alignments to the green anole - bearded dragon genome alignment (0.26% of the bearded dragon genome) and 14.5 Mb to the alligator - turtle alignment (0.61% of the turtle genome) (Figure 3, Supplementary Table 2). Thus, despite the fact that reptile and mammal genomes generally have a similar repeat content of ~30-50% [28, 30], RepeatFiller added fewer alignment for reptiles compared with mammals. This shows that other factors in addition to genomic repeat content also

influence the amount of added alignments. Nevertheless, more than one megabase of previously-undetected alignments for birds or reptiles show that RepeatFiller, with little additional runtime, can also improve the completeness of aligning repetitive regions between species in these groups.

RepeatFiller application uncovers thousands of novel repeat-derived conserved non-exonic elements

Next, we investigated whether some of the newly-aligning sequences show evidence of evolutionary constraint, which indicates purifying selection and a biological function. To this end, we used the pairwise alignments, generated either with or without RepeatFiller, to build two human-referenced multiple genome alignments of 21 mammals with Multiz [31]. Then, we used PhastCons [32] to identify constrained elements. We found that the majority (98%) of the 164 Mb in the human genome that are classified as constrained in the multiple alignment without RepeatFiller were also classified as constrained in the RepeatFiller-subjected alignment.

Dividing the conserved regions detected in the alignment without RepeatFiller into exonic and non-exonic regions, we found that 99.8% of the exonic and 97.4% of the non-exonic regions are also classified as constrained in the RepeatFiller-subjected alignment. Since conserved exonic regions are virtually identical, likely because they rarely overlap repeats, we focused our comparison on the conserved non-exonic elements (CNEs), which often overlap *cis*-regulatory elements [33-35]. This comparison first showed that 3.46 Mb of the human genome were newly classified as conserved non-exonic in the RepeatFiller-subjected alignment, representing 2.9% of all conserved non-exonic bases detected in this alignment. Requiring a minimum size of 30 bp, application of RepeatFiller led to the identification of 30167 novel CNEs that are listed in Supplementary Table 3. With a median size of 41 bp, these novel CNEs are shorter than CNEs already detected in the non-RepeatFiller alignments (median 50 bp, two-sided Wilcoxon rank sum test $P<e^{-16}$, Supplementary Figure 1), likely because most of the longer conserved regions were already in the initial genome-wide alignment step. Consistent with previous findings that CNEs are in general more AT-rich [36], we found that the novel CNEs are more AT-rich than randomly selected, non-conserved genomic regions (two-sided Wilcoxon rank sum test $P<e^{-16}$, Supplementary Figure 1).

Two striking examples of newly-identified CNEs are shown in Figures 4 and 5. Figure 4 shows the genomic region overlapping *MEIS3*, a homeobox transcription factor gene that synergizes with Hox genes and is required for hindbrain development and survival of pancreatic beta-cells [37-39]. By revealing novel alignments to many non-human

mammals, RepeatFiller identifies several novel repeat-overlapping CNEs in introns of *MEIS3* (Figure 4). Figure 5 shows the genomic region around *AUTS2*, a transcriptional regulator required for neurodevelopment that is associated with human neurological disorders such as autism [40, 41]. Applying RepeatFiller revealed several novel CNEs upstream of *AUTS2*. For some of these CNEs, RepeatFiller incorporated a well-aligning sequence of 19 mammals, which then permitted the identification of evolutionary constraint. Overall, applying RepeatFiller led the identification of more than 30000 CNEs that were not detected before.

RepeatFiller improves annotations of Conserved Non-exonic Elements

Interestingly, the comparison of conserved non-exonic bases detected by PhastCons also revealed 3.08 Mb of the human genome that were classified as conserved non-exonic only in the multiple alignment without RepeatFiller, but not in the RepeatFiller-subjected alignment. These 3.08 Mb represent 2.6% of all conserved non-exonic bases detected in the alignment without RepeatFiller. The 29334 CNEs with a size ≥30 bp are listed in Supplementary Table 4. To investigate the reasons underlying these 'lost' CNEs, we first sought to confirm that the RepeatFiller-subjected alignment had an increased species coverage in these regions. Indeed, we found that RepeatFiller added on average 3.9 (median 3) aligning species to these lost CNEs. Inspecting many of these CNEs showed that the newly added sequences are similar to the already-aligned sequences; however, they exhibit more substitutions. These substitutions increase the overall sequence divergence across mammals, which likely explains why the same region was not classified as constrained anymore, despite having a higher coverage of aligning species. Figure 6A and B shows two examples of such genomic regions that are not classified as constrained after adding additional alignments with RepeatFiller.

To confirm that the newly-added sequences increase the overall sequence divergence, we applied GERP++ [42] to both multiple alignments (Supplementary Figure 2A). For each alignment column, GERP++ estimates the number of substitutions that were rejected by purifying selection (RS = rejected substitutions) by subtracting the number of observed substitutions from the number of substitutions expected under neutrality. Since GERP++ computes the number of substitutions expected under neutrality from a phylogenetic tree that is pruned to the aligning species (Supplementary Figure 2B), we can directly compare RS between alignment columns that were only classified as constrained in either alignment to estimate whether the RepeatFiller-added sequences evolve slower than expected under neutrality. Specifically, for each alignment column, we computed the difference in RS before and after adding new alignments with RepeatFiller, as illustrated in Supplementary Figure 2B.

We found that the alignment columns, where constraint was only detected in the alignment without RepeatFiller, mostly exhibit slightly negative RS differences (Figure 6C, grey background), which suggests that many positions in the RepeatFiller-added sequences do not evolve under strong constraint. Hence, the extent of constraint in the more limited set of aligning sequences was likely overestimated, providing an explanation of why these genomic regions were not classified anymore as constrained across placental mammals. It should be noted that these regions may still be under constraint in particular lineages. In contrast, most alignment columns, where constraint was only detected after applying RepeatFiller, exhibit a positive RS difference (Figure 6C, orange background), which suggests that the newly-added sequences evolve under constraint. Overall, by uncovering previously-unknown alignments, RepeatFiller application led to an improved CNE annotation.

Discussion

While transposon-derived sequences can be co-opted into a multitude of biological roles and can evolve under evolutionary constraint, comprehensively detecting alignments between ancestral transposons and other repeats is not straightforward. The main reason is that considering all repeat-overlapping alignment seeds during the initial whole genome alignment step is computationally not feasible. However, it is feasible to consider all seeds when aligning local regions that are bounded by colinear aligning blocks. We provide a tool RepeatFiller that implements this idea and incorporates newly-detected repeat-overlapping alignments into pairwise alignment chains. We tested the tool on alignments between human and 20 representative mammals and showed that with little additional computational runtime RepeatFiller uncovers between 22 and 84 Mb of previously-undetected alignments that mostly originate from transposable elements. We also showed that RepeatFiller can detect megabases of previously-undetected alignments for fragmented mammalian genomes or for genomes of birds and reptiles, suggesting that RepeatFiller can be applied to genome alignments of a wide range of species.

We further show that RepeatFiller application enables a refined and more complete CNE annotation by two means. First, applying RepeatFiller led the identification of thousands of CNEs whose aligning sequences were not detected before. This includes highly-conserved transposon-derived CNEs that are located near important developmental genes. Second, the sequences added by RepeatFiller may not evolve slower than expected under neutral evolution. In this case, providing a more complete set of aligning sequences led to the removal of thousands of putatively-spurious CNEs that overall do

not evolve under strong constraint across placental mammals, though the possibility of lineage-specific constraint remains.

Taken together, RepeatFiller implements an efficient way to improve the completeness of aligning repetitive regions in whole genome alignments, which helps annotating conserved non-exonic elements and studying transposon co-option and genome evolution.

Materials and Methods

Generating pairwise genome alignments

For all mammalian species, we used the human hg38 genome assembly as the reference genome. For the alignments of non-mammalian species, the reference assemblies are specified in Supplementary Table 2. To compute pairwise genome alignments, we used lastz version 1.04.00 [21] and the chain/net pipeline [26] with default parameters (chainMinScore 1000, chainLinearGap loose). We used the lastz alignment parameters K = 2400, L = 3000, Y = 9400, H = 2000 and the lastz default scoring matrix. We also tested aligning human and rhesus macaque using K = 4500, L = 3000, Y = 15000, H = 2000 and the UCSC human_chimp.v2.q scoring matrix and found that applying RepeatFiller to these chains also added a similar amount (25.2 vs. 22.4 Mb) of newly aligning sequence (Supplementary Table 1). All species names and their assemblies are listed in Supplementary Tables 1 and 2.

RepeatFiller

The input of RepeatFiller is a file containing co-linear chains of local alignment blocks. This file must be in the UCSC chain format as defined here https://genome.ucsc.edu/goldenPath/help/chain.html. The output is a file that contains the same chains plus the newly-added local alignment blocks. By default, RepeatFiller only considers unaligned regions in both the reference and query genome that are at least 30 bp and at most 20000 bp long. We considered all chains with the score greater than 25000. For each unaligning region that fulfills the size thresholds, RepeatFiller uses lastz with the same parameters as above but with a slightly more sensitive ungapped alignment threshold (K=2000). All repeat-masking (lower case letters) was removed before providing the local sequences to lastz. Since lastz may find multiple additional local alignments in this second step, we used axtChain [26] to obtain a ‘mini chain’ of local alignments for this unaligning region. RepeatFiller then inserts the aligning blocks of a newly-detected mini chain at the respective position in the original chain if the score of the mini chain is at least 5000. All default parameters for the size of unaligning regions,

minimum chain scores and local alignment parameters can be changed by the user via parameters. Finally, RepeatFiller recomputes the score of the entire chain if new alignments were added.

We compared the number of aligning bases in the chains before and after applying RepeatFiller. To this end, we used the coordinates of aligning chain blocks to determine how many bases of the human hg38 assembly align (via at least one chain) to the query species. We used the RepeatMasker repeat annotation for hg38, available at the UCSC Genome Browser [43], to determine how many of the newly-added alignments overlap repetitive elements.

Generating multiple alignments

Before building multiple alignment, we filtered out low scoring chains and nets requiring a minimum score of 100000. We used Multiz-tba [31] with default parameters to generate two reference-based multiple alignments using the pairwise alignment nets produced with and without RepeatFiller, respectively.

Conservation analysis

To identify constrained elements, one needs a tree with branch lengths representing the number of substitutions per neutral site. We used four-fold degenerated codon sites based on the human ENSEMBL gene annotation to estimate the neutral branch lengths with PhyloFit [32]. To identify conserved regions, we used PhastCons [32] with the following parameters: rho=0.31; expected-length=45; target-coverage=0.3. To obtain conserved non-exonic regions, we first obtained exonic regions from the human Ensembl and RefSeq annotation (UCSC tables ensGene and refGene). As done before [25], we merged all exonic regions and added 50 bp flanks to exclude splice site proximal regions that often harbor conserved splicing regulatory elements. To obtain Conserved Non-exonic Elements (CNEs), we subtracted these exonic bases and their flanks from all conserved regions.

To compare constraint in genomic regions classified as constraint in only one alignment, we used GERP++ (RRID:SCR_000563)[42] with default parameters (acceptable false positive rate = 0.05) to estimate constraint per genomic position. We denote genomic regions as ‘gained’ if they were classified as constrained by PhastCons only in the multiple alignment generated with RepeatFiller. We denote genomic regions as ‘lost’ if they were classified as constrained only in alignment generated without RepeatFiller (Supplementary Figure 2A). Gained and lost regions were identified using ‘bedtools intersect’ (RRID:SCR_006646)[44]. For each position in ‘gained’ and ‘lost’ non-exonic regions, we computed the RS score (number of rejected substitutions) with GERP++ [42]

and calculated the difference between the RS score obtained for the alignment with and without RepeatFiller (Supplementary Figure 2B). These differences are plotted in Figure 6C. Positive differences indicate that the sequences added by RepeatFiller evolve slower than under neutrality, thus increasing the number of rejected substitutions. Differences close to zero indicate that the newly-added sequences evolve as expected under neutral evolution and negative differences indicate that they evolve faster than expected under neutral evolution.

Availability of supporting source code and requirements

- Project name: RepeatFiller
- Project home page: https://github.com/hillerlab/GenomeAlignmentTools
- Programming language: perl and python
- Other requirements: lastz
- License: MIT License
- RRID: SCR_017414
- ELIXIR bio.tools registry: biotools:RepeatFiller

Data Availability

The multiple genome alignments generated with and without applying RepeatFiller and the respective PhastCons conserved elements are available at https://bds.mpi-cbg.de/hillerlab/RepeatFiller/. The CNEs that differ between both alignments are available in Supplementary Tables 3 and 4. The RepeatFiller source code is available at https://github.com/hillerlab/GenomeAlignmentTools. Other supporting data and code snapshots are available from the *GigaScience* GigaDB repository[45].

Abbreviations

bp: base pairs; CNE: Conserved Non-exonic Element; CPU: central processing unit; Kb: kilo basepair; Mb: mega base-pair; RS: rejected substitutions.

Competing interests

The authors have no competing interests.

Acknowledgment

We thank the genomics community for sequencing and assembling the genomes and the UCSC genome browser group for providing software and genome annotations. We also thank the Computer Service Facilities of the MPI-CBG and MPI-PKS for their support.

Funding

This work was supported by the Max Planck Society and the Leibniz Association (SAW-2016-SGN-2).

References

1. Ivancevic AM, Kortschak RD, Bertozzi T and Adelson DL. LINEs between Species: Evolutionary Dynamics of LINE-1 Retrotransposons across the Eukaryotic Tree of Life. Genome Biol Evol. 2016;8 11:3301-22. doi:10.1093/gbe/evw243.
2. Sotero-Caio CG, Platt RN, 2nd, Suh A and Ray DA. Evolution and Diversity of Transposable Elements in Vertebrate Genomes. Genome Biol Evol. 2017;9 1:161-77. doi:10.1093/gbe/evw264.
3. Meader S, Ponting CP and Lunter G. Massive turnover of functional sequence in human and other mammalian genomes. Genome Res. 2010;20 10:1335-43. doi:10.1101/gr.108795.110.
4. Feschotte C. Transposable elements and the evolution of regulatory networks. Nature reviews Genetics. 2008;9 5:397-405. doi:10.1038/nrg2337.
5. Chuong EB, Elde NC and Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. Nature reviews Genetics. 2017;18 2:71-86. doi:10.1038/nrg.2016.139.
6. Sorek R, Ast G and Graur D. Alu-containing exons are alternatively spliced. Genome Res. 2002;12 7:1060-7. doi:10.1101/gr.229302.
7. Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, et al. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature. 2006;441 7089:87-90. doi:10.1038/nature04696.
8. Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet. 2010;42 7:631-4. doi:10.1038/ng.600.
9. Lynch VJ, Leclerc RD, May G and Wagner GP. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. Nat Genet. 2011;43 11:1154-9. doi:10.1038/ng.917.
10. Batut P, Dobin A, Plessy C, Carninci P and Gingeras TR. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. Genome Res. 2013;23 1:169-80. doi:10.1101/gr.139618.112.
11. Chuong EB, Rumi MA, Soares MJ and Baker JC. Endogenous retroviruses function as species-specific enhancer elements in the placenta. Nat Genet. 2013;45 3:325-9. doi:10.1038/ng.2553.
12. Notwell JH, Chung T, Heavner W and Bejerano G. A family of transposable elements co-opted into developmental enhancers in the mouse neocortex. Nature communications. 2015;6:6644. doi:10.1038/ncomms7644.
13. Chuong EB, Elde NC and Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science. 2016;351 6277:1083-7. doi:10.1126/science.aad5497.
14. Rech GE, Bogaerts-Marquez M, Barron MG, Merenciano M, Villanueva-Canas JL, Horvath V, et al. Stress response, behavior, and development are shaped by transposable element-induced mutations in Drosophila. PLoS Genet. 2019;15 2:e1007900. doi:10.1371/journal.pgen.1007900.

15. Villanueva-Canas JL, Horvath V, Aguilera L and Gonzalez J. Diverse families of transposable elements affect the transcriptional regulation of stress-response genes in Drosophila melanogaster. Nucleic Acids Res. 2019; doi:10.1093/nar/gkz490.
16. Lowe CB, Bejerano G and Haussler D. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. Proceedings of the National Academy of Sciences of the United States of America. 2007;104 19:8005-10.
17. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011;478 7370:476-82. doi:10.1038/nature10530.
18. Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990;215 3:403-10. doi:10.1016/S0022-2836(05)80360-2.
19. Smit A, Hubley R and Green P: RepeatMasker Open-4.0. http://www.repeatmasker.org (2013-2015).
20. Morgulis A, Gertz EM, Schaffer AA and Agarwala R. WindowMasker: window-based masker for sequenced genomes. Bioinformatics. 2006;22 2:134-41. doi:10.1093/bioinformatics/bti774.
21. Harris RS. *Improved pairwise alignment of genomic DNA*. The Pennsylvania State University, 2007.
22. Zeng L, Kortschak RD, Raison JM, Bertozzi T and Adelson DL. Superior ab initio identification, annotation and characterisation of TEs and segmental duplications from genome assemblies. PloS one. 2018;13 3:e0193588. doi:10.1371/journal.pone.0193588.
23. Kielbasa SM, Wan R, Sato K, Horton P and Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Res. 2011;21 3:487-93. doi:10.1101/gr.113985.110.
24. Sharma V and Hiller M. Increased alignment sensitivity improves the usage of genome alignments for comparative gene annotation. Nucleic Acids Res. 2017;45 14:8369-77. doi:10.1093/nar/gkx554.
25. Hiller M, Agarwal S, Notwell JH, Parikh R, Guturu H, Wenger AM, et al. Computational methods to detect conserved non-genic elements in phylogenetically isolated genomes: application to zebrafish. Nucleic Acids Res. 2013;41 15:e151. doi:10.1093/nar/gkt557.
26. Kent WJ, Baertsch R, Hinrichs A, Miller W and Haussler D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proceedings of the National Academy of Sciences of the United States of America. 2003;100 20:11484-9. doi:10.1073/pnas.1932072100.
27. Warren WC, Hillier LW, Tomlinson C, Minx P, Kremitzki M, Graves T, et al. A New Chicken Genome Assembly Provides Insight into Avian Genome Structure. G3 (Bethesda). 2017;7 1:109-17. doi:10.1534/g3.116.035923.
28. Kapusta A, Suh A and Feschotte C. Dynamics of genome size evolution in birds and mammals. Proceedings of the National Academy of Sciences of the United States of America. 2017;114 8:E1460-E9. doi:10.1073/pnas.1616702114.
29. Drosophila 12 Genomes C, Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, et al. Evolution of genes and genomes on the Drosophila phylogeny. Nature. 2007;450 7167:203-18. doi:10.1038/nature06341.
30. Roscito JG, Sameith K, Pippel M, Francoijs KJ, Winkler S, Dahl A, et al. The genome of the tegu lizard Salvator merianae: combining Illumina, PacBio, and optical mapping data to generate a highly contiguous assembly. Gigascience. 2018; doi:10.1093/gigascience/giy141.
31. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, et al. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. 2004;14 4:708-15. doi:10.1101/gr.1933104.

32. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15 8:1034-50.
33. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, et al. Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol. 2005;3 1:e7.
34. Visel A, Prabhakar S, Akiyama JA, Shoukry M, Lewis KD, Holt A, et al. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. Nat Genet. 2008;40 2:158-60. doi:10.1038/ng.2007.55.
35. Wittkopp PJ and Kalay G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nature reviews Genetics. 2011;13 1:59-69. doi:10.1038/nrg3095.
36. Polychronopoulos D, King JWD, Nash AJ, Tan G and Lenhard B. Conserved non-coding elements: developmental gene regulation meets genome organization. Nucleic Acids Res. 2017;45 22:12611-24. doi:10.1093/nar/gkx1074.
37. Vlachakis N, Choe SK and Sagerstrom CG. Meis3 synergizes with Pbx4 and Hoxb1b in promoting hindbrain fates in the zebrafish. Development. 2001;128 8:1299-312.
38. Dibner C, Elias S and Frank D. XMeis3 protein activity is required for proper hindbrain patterning in Xenopus laevis embryos. Development. 2001;128 18:3415-26.
39. Liu J, Wang Y, Birnbaum MJ and Stoffers DA. Three-amino-acid-loop-extension homeodomain factor Meis3 regulates cell survival via PDK1. Proceedings of the National Academy of Sciences of the United States of America. 2010;107 47:20494-9. doi:10.1073/pnas.1007001107.
40. Gao Z, Lee P, Stafford JM, von Schimmelmann M, Schaefer A and Reinberg D. An AUTS2-Polycomb complex activates gene expression in the CNS. Nature. 2014;516 7531:349-54. doi:10.1038/nature13921.
41. Amarillo IE, Li WL, Li X, Vilain E and Kantarci S. De novo single exon deletion of AUTS2 in a patient with speech and language disorder: a review of disrupted AUTS2 and further evidence for its role in neurodevelopmental disorders. Am J Med Genet A. 2014;164A 4:958-65. doi:10.1002/ajmg.a.36393.
42. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A and Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS computational biology. 2010;6 12:e1001025. doi:10.1371/journal.pcbi.1001025.
43. Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, et al. The UCSC Genome Browser database: 2019 update. Nucleic Acids Res. 2019;47 D1:D853-D8. doi:10.1093/nar/gky1095.
44. Quinlan AR and Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26 6:841-2. doi:10.1093/bioinformatics/btq033.
45. Osipova E; Hecker N; Hiller M (2019): Supporting data for "RepeatFiller newly identifies megabases of aligning repetitive sequences and improves annotations of conserved non-exonic elements" GigaScience Database. http://dx.doi.org/10.5524/100656

Figures

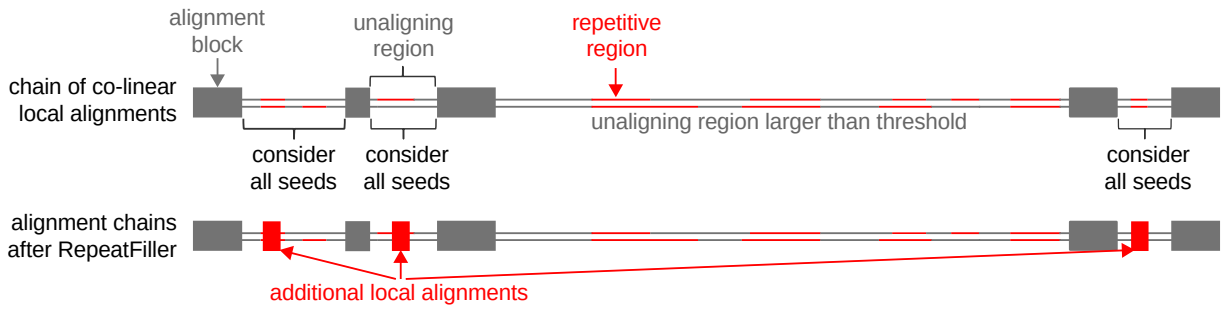


Figure 1: Missed repeat-overlapping alignments and concept of RepeatFiller. Illustration of RepeatFiller. Focusing on unaligning regions in a reference and query genome that are flanked by up- and downstream aligning blocks, RepeatFiller performs a second round of local alignment considering also repeat-overlapping seeds. Newly found local alignments (red boxes) are inserted into the context of other aligning blocks (grey boxes). Unaligning regions that are larger than a user-defined threshold are not considered as the chance of aligning non-orthologous repeats is increased.

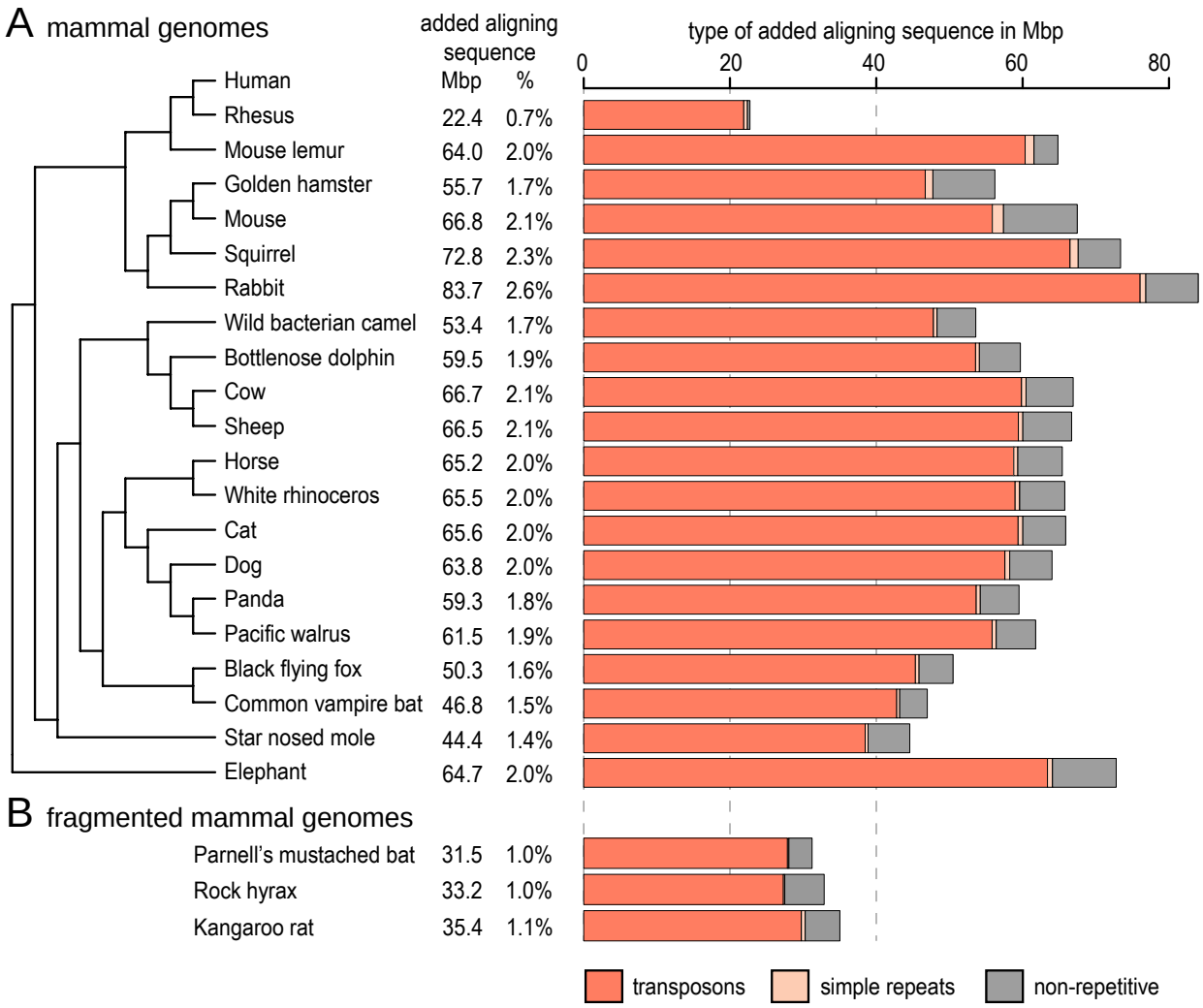


Figure 2: RepeatFiller adds several megabases of aligning transposable elements to existing mammalian genome alignments.

(A) Phylogenetic tree of human and 20 non-human mammals whose genomes we aligned to the human genome. The amount of newly alignments detected by RepeatFiller is shown in megabases and in percent relative to the human genome. Bar charts provide a breakdown of newly-added aligning sequences into overlap with transposons, simple repeats and non-repetitive sequence.

(B) Application of RepeatFiller to fragmented mammalian assemblies still adds a substantial amount of new alignments.

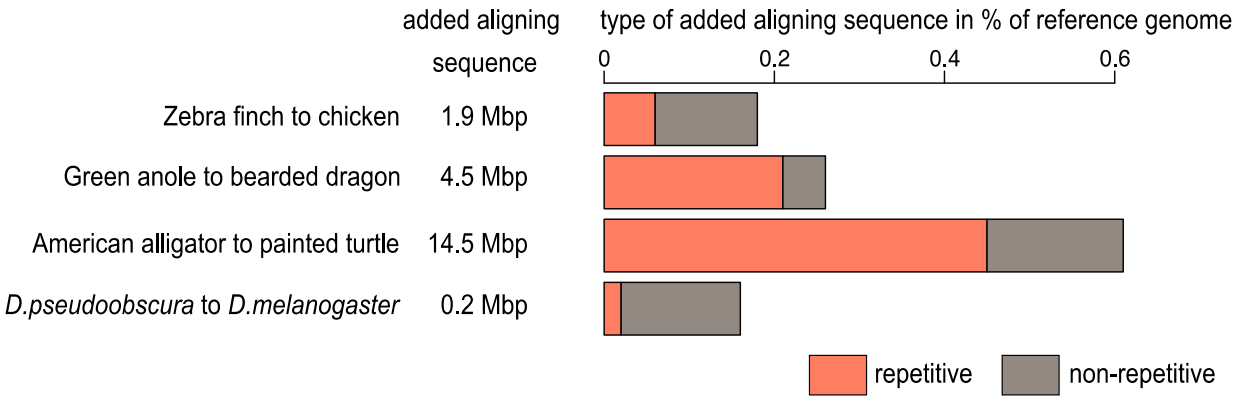


Figure 3: RepeatFiller also detects additional alignments for non-mammalian genomes. The figure shows how many new alignments were detected by applying RepeatFiller to pairwise alignments of birds, reptiles and *Drosophilids*. Both the amount (in megabases) of new alignments and the percent of the reference genome additionally aligned are shown. Bar charts show which portion of newly-added alignments overlap repetitive sequences.

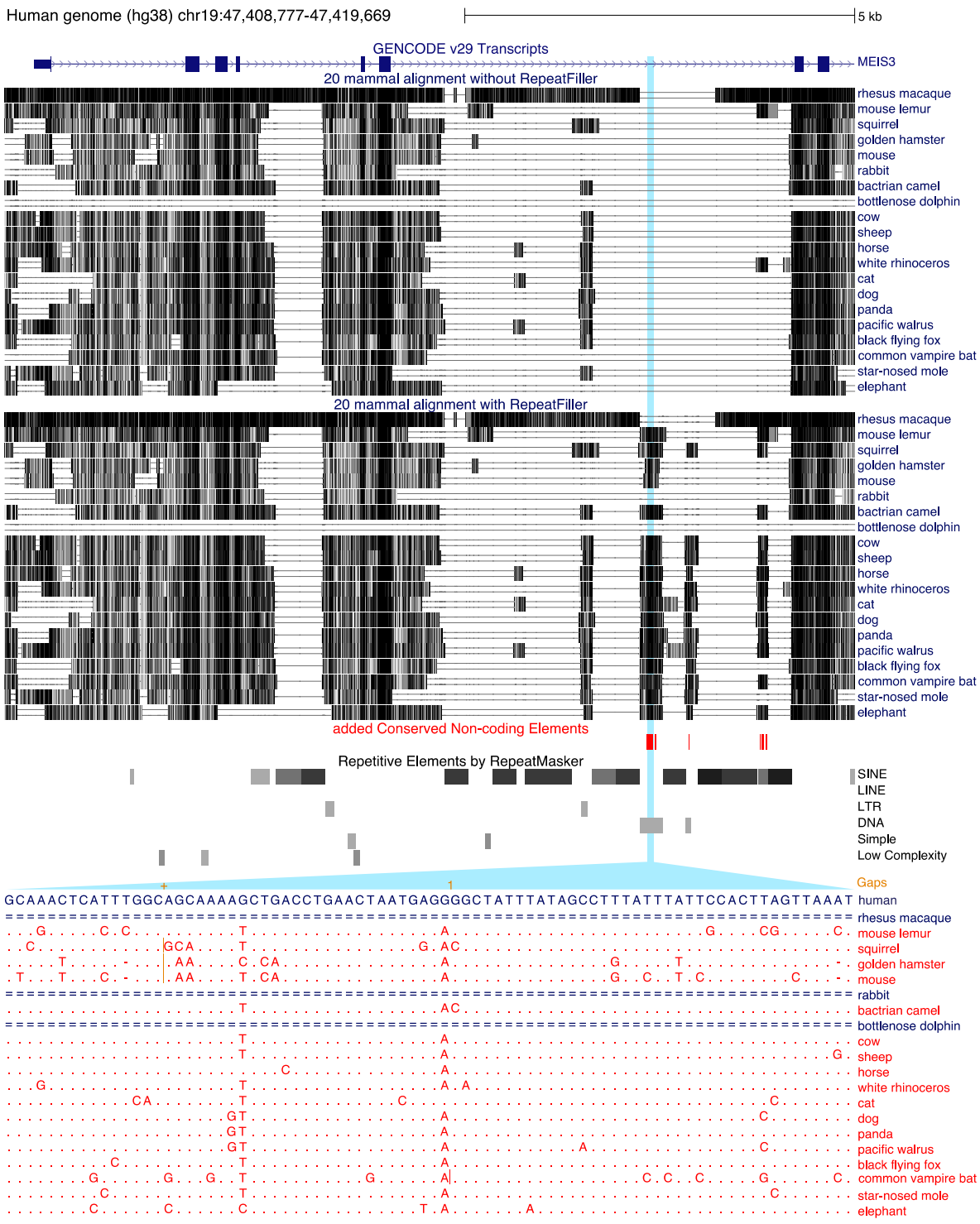


Figure 4: Examples of newly-identified CNEs near *MEIS3*.

UCSC genome browser [43] screenshot shows an ~11 kb genomic region overlapping the gene *MEIS3*, a homeobox transcription factor that is required for hindbrain development. Visualization of the two multiple genome alignments (without RepeatFiller at the top, with RepeatFiller below; boxes representing align regions with darker colors

indicating a higher alignment identity) shows that RepeatFiller adds several aligning sequences, some of which evolve under evolutionary constraint and thus are CNEs (red boxes) only detected in the RepeatFiller-subjected alignment. The RepeatMasker annotation shows that these newly-identified CNEs overlap transposons. The zoom-in shows the 21-mammal alignment of one of the newly-identified CNEs, which overlaps a DNA transposon. While this genomic region did not align to any mammal before applying RepeatFiller, our tool identified a well-aligning sequence for 17 non-human mammals (red font). A dot represents a base that is identical to the human base, insertions are marked by vertical orange lines, and unaligning regions are showed as double lines.

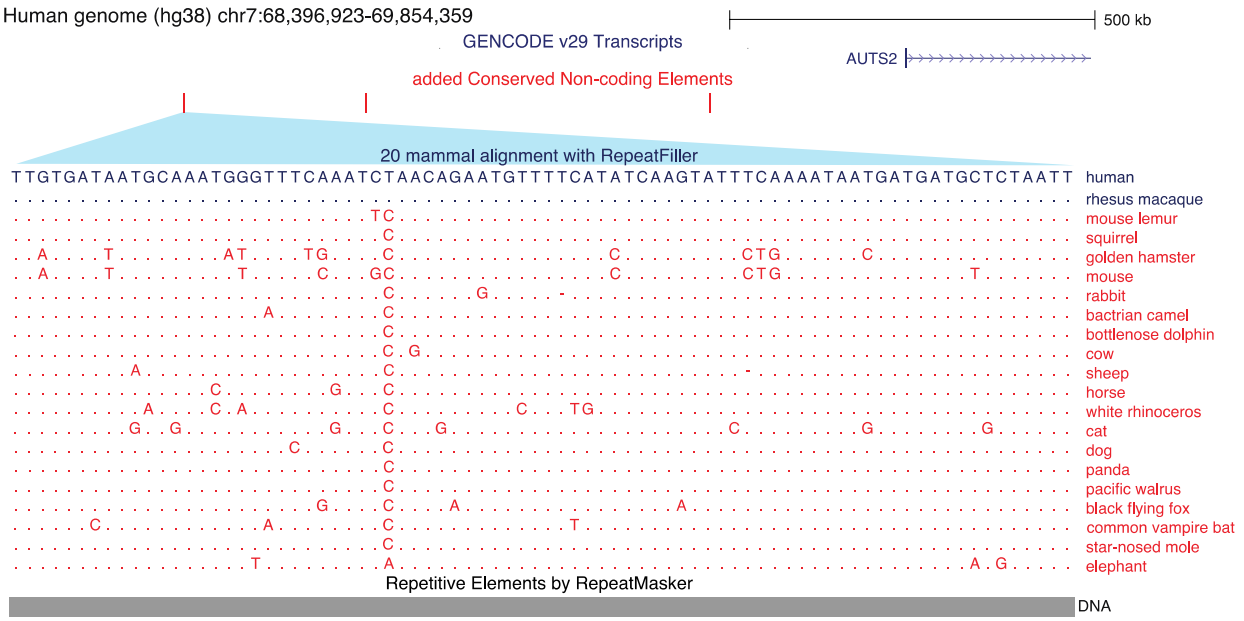


Figure 5: Examples of newly-identified CNEs upstream of *AUTS2*.
UCSC genome browser screenshot shows a ~1.5 Mb genomic region around *AUTS2*, a transcriptional regulator required for neurodevelopment. CNEs only detected in the RepeatFiller-subjected multiple alignment are marked as red tick marks. The zoom-in shows the 21-mammal alignment of one of the newly-identified CNEs. While only the rhesus macaque sequence aligned to human before applying RepeatFiller, our tool identifies a well-aligning sequence for all 19 other mammals (red font). A dot represents a base that is identical to the human base. The RepeatMasker annotation (bottom) shows that this newly-identified CNE overlaps a DNA transposon.

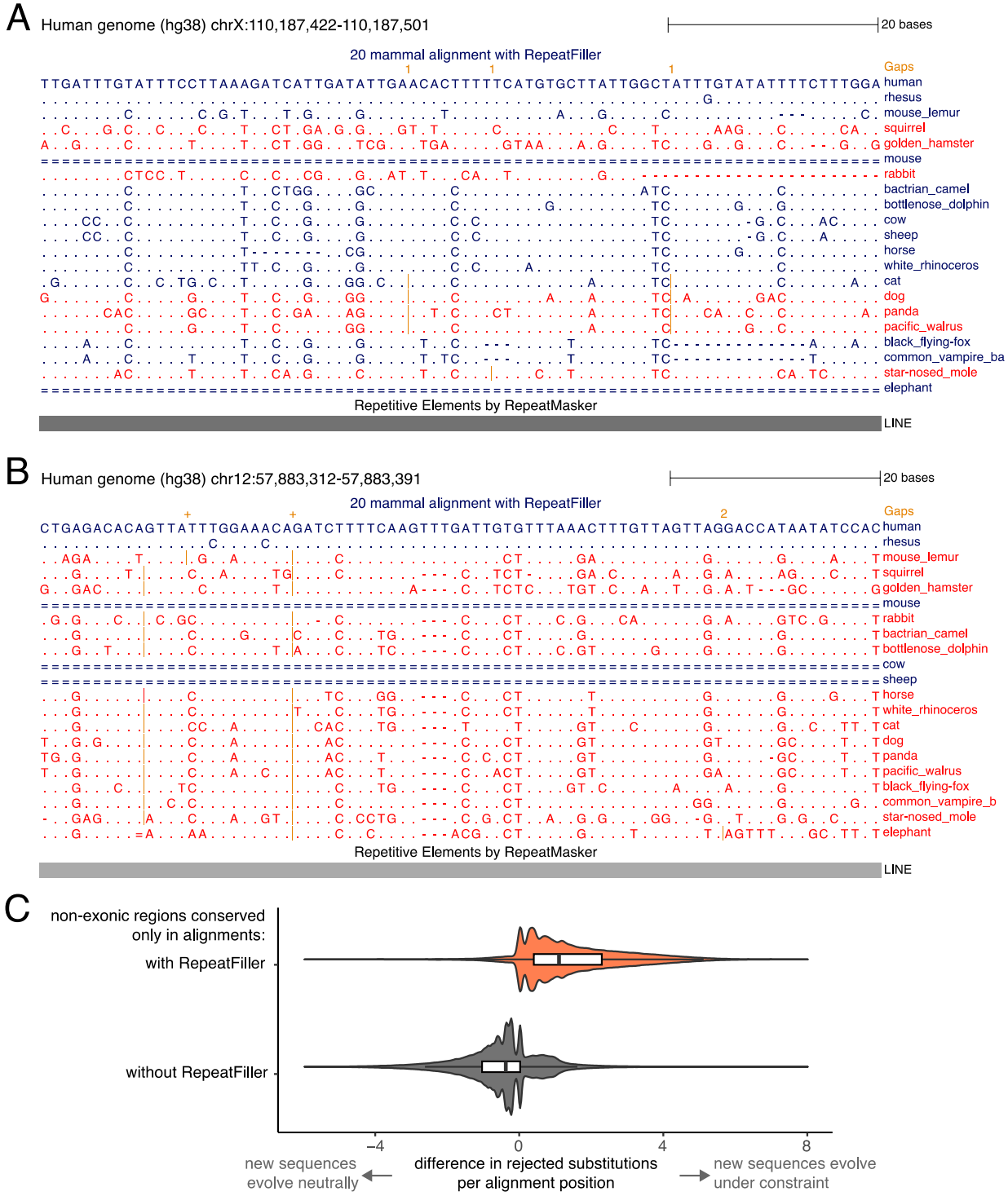


Figure 6: Additional alignments found with RepeatFiller reveal absence of conservation in the genomic regions that were erroneously classified as conserved before.

(A, B) UCSC genome browser screenshots showing two examples of genomic regions that were only classified as constrained in a multiple genome alignment generated without

applying RepeatFiller. Dots in these alignments represent bases that are identical to the human base, insertions are marked by vertical orange lines, and unaligning regions are showed as double lines. The alignments show that the sequences of species added by RepeatFiller (red font) exhibit a number of substitutions. This explains why these regions were not classified as constrained anymore, despite adding more aligning sequences. Please note that in (B) only the sequence of the rhesus macaque was aligned before applying RepeatFiller. Sequences in both (A) and (B) overlap LINE transposons.
(C) Difference in evolutionary constraint in non-exonic alignment columns that are only classified as constrained in either alignment. For each alignment position, we used GERP++ to compute the estimated number of substitutions rejected by purifying selection (RS). The difference in RS between alignments with and without RepeatFiller is visualized as a violin plot overlaid with a white box plot. This shows that almost all non-exonic bases that were only detected as constrained in the alignment with RepeatFiller (orange background) have a positive RS difference, indicating that the newly-aligning sequences added by RepeatFiller largely evolve under evolutionary constraint. In contrast, non-exonic bases only detected as constrained in the alignment without RepeatFiller (grey background) often have slightly negative RS differences, indicating that many of the newly-added sequences do not evolve under constraint. The two distributions are significantly different ($P<e^{-16}$, two-sided Wilcoxon rank sum test).

Figure 1

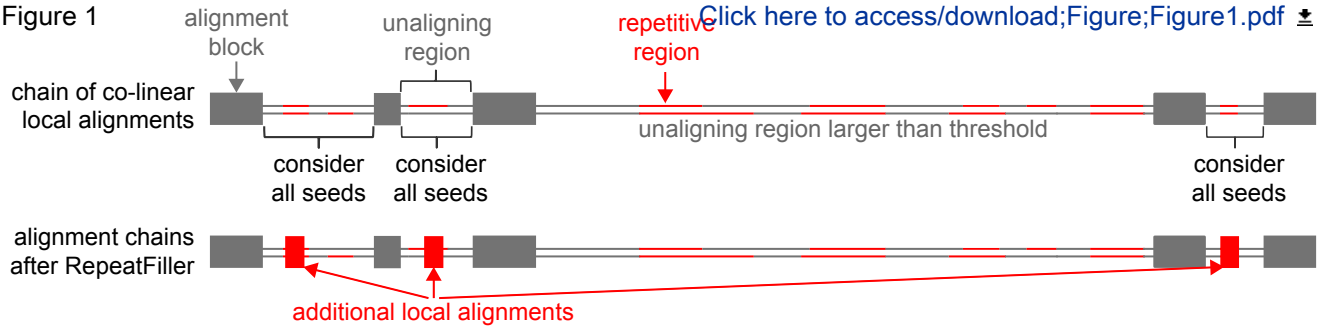


Figure 2

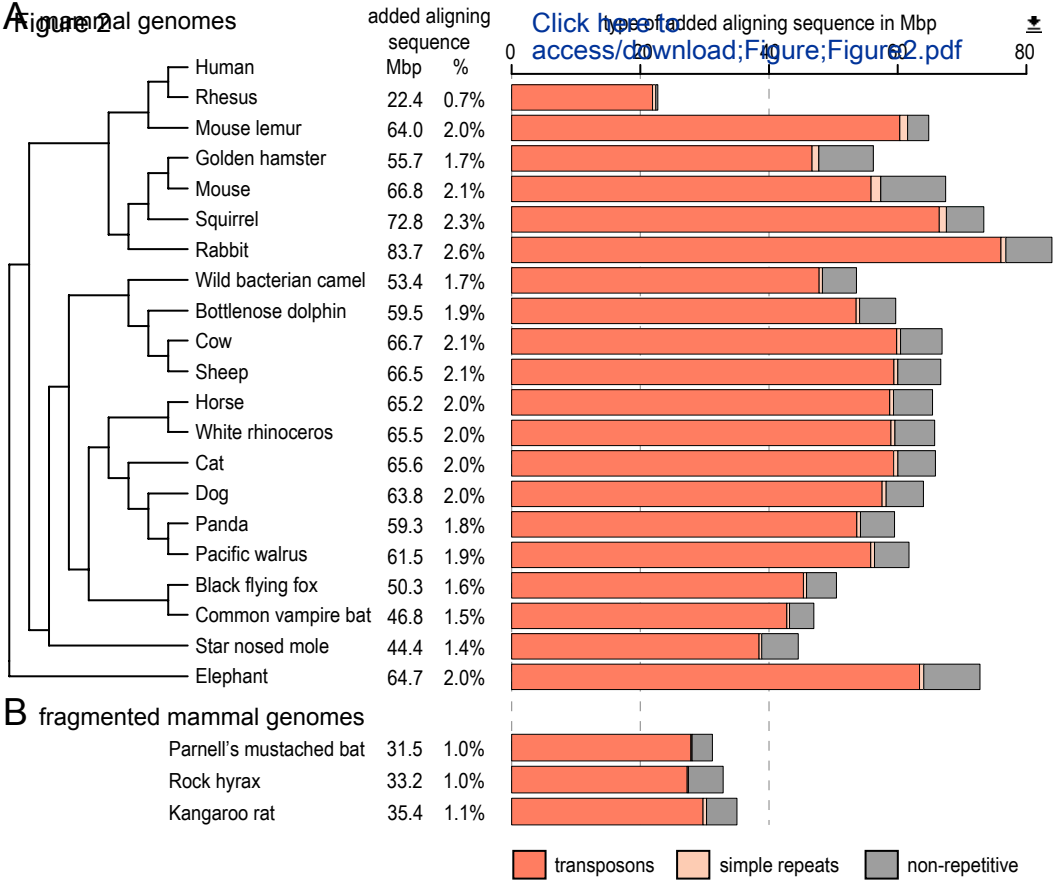


Figure 3

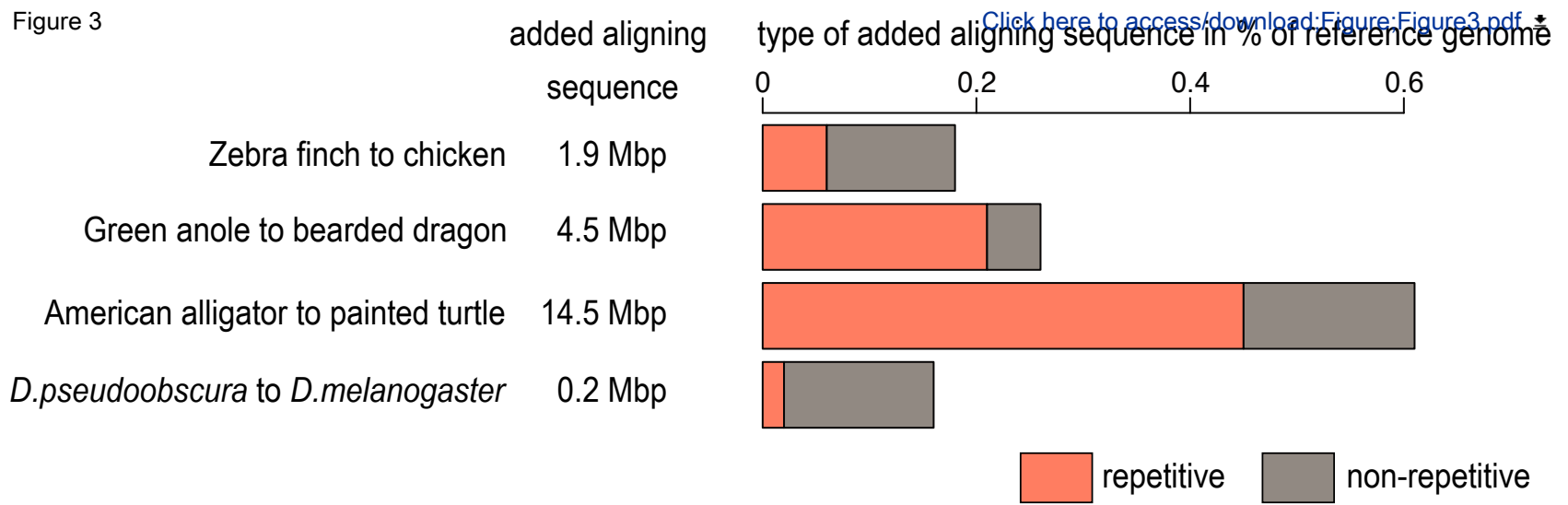


Figure 4

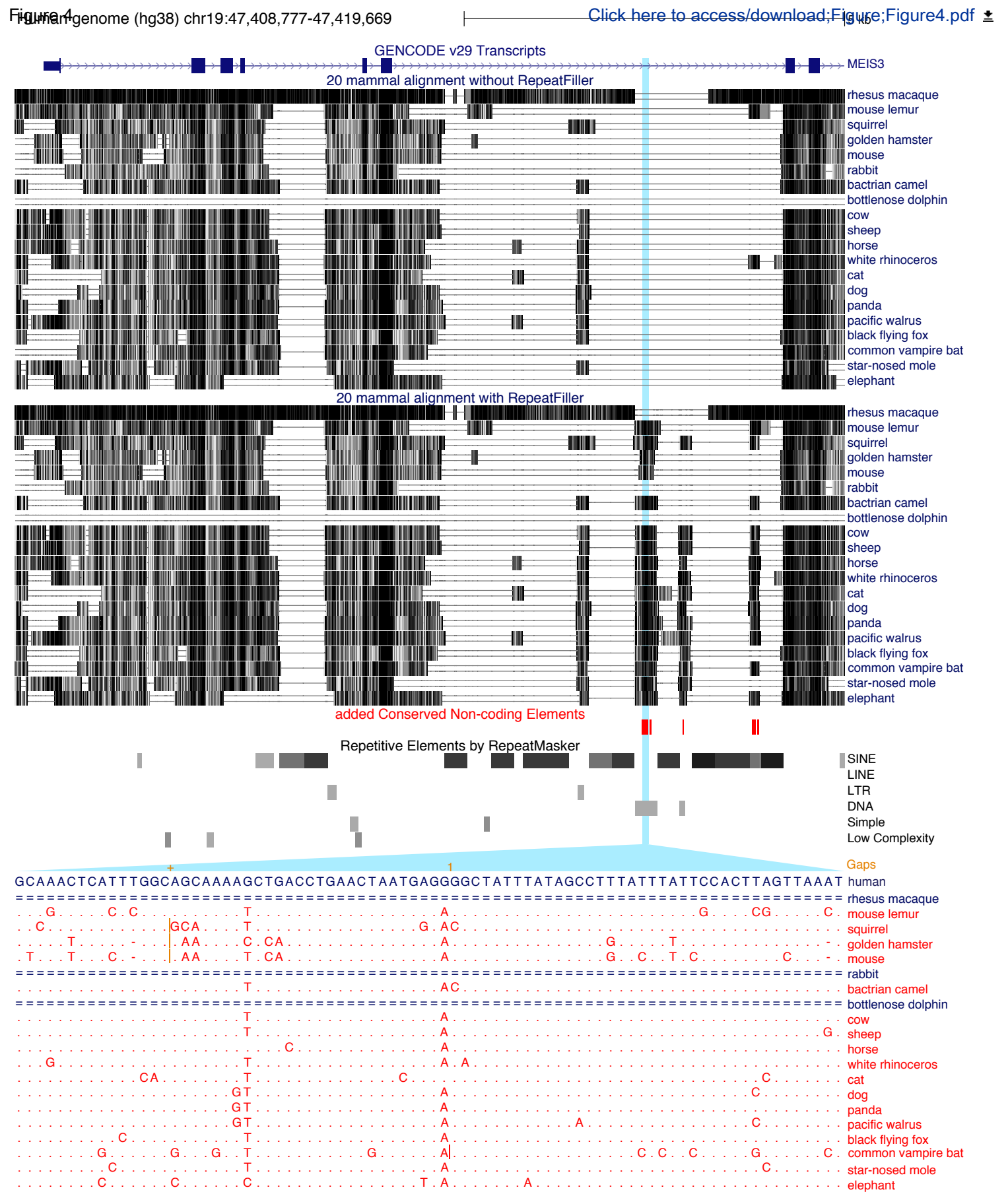

Human genome (hg38) chr19:47,408,777-47,419,669
5 kb
Click here to access/download;Figure;Figure4.pdf
GENCODE v29 Transcripts
MEIS3
20 mammal alignment without RepeatFiller
rhesus macaque
mouse lemur
squirrel
golden hamster
mouse
rabbit
bactrian camel
bottlenose dolphin
cow
sheep
horse
white rhinoceros
cat
dog
panda
pacific walrus
black flying fox
common vampire bat
star-nosed mole
elephant
20 mammal alignment with RepeatFiller
rhesus macaque
mouse lemur
squirrel
golden hamster
mouse
rabbit
bactrian camel
bottlenose dolphin
cow
sheep
horse
white rhinoceros
cat
dog
panda
pacific walrus
black flying fox
common vampire bat
star-nosed mole
elephant
added Conserved Non-coding Elements
Repetitive Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Gaps
GCAAACTCATTTGGCAGCAAAAGCTGACCTGAACTAATGAGGGGCTATTTATAGCCTTTATTTATTCCACTTAGTTAAAT human
rhesus macaque
mouse lemur
squirrel
golden hamster
mouse
rabbit
bactrian camel
bottlenose dolphin
cow
sheep
horse
white rhinoceros
cat
dog
panda
pacific walrus
black flying fox
common vampire bat
star-nosed mole
elephant

Figure 5

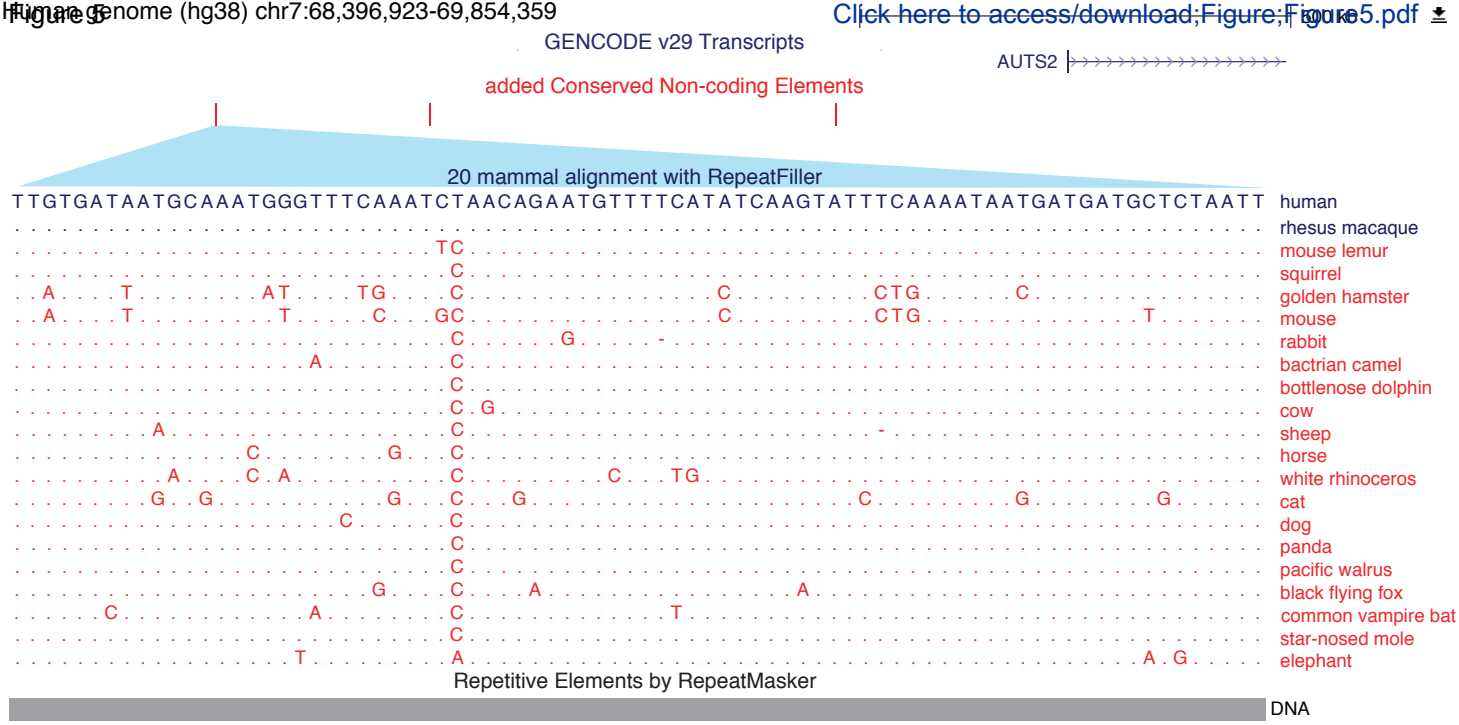
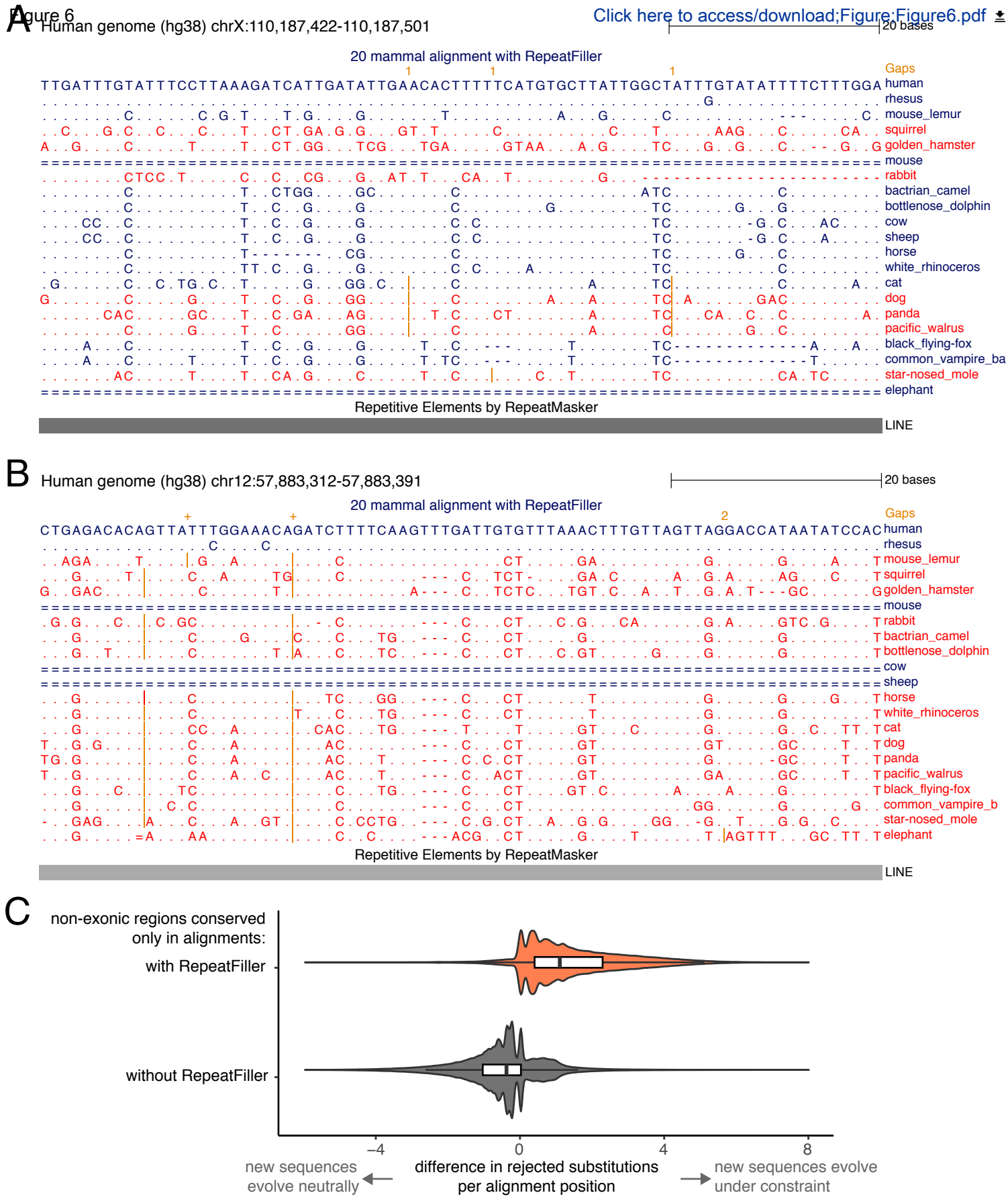


Figure 6

Click here to access/download;Figure;Figure6.pdf



Supplementary Figures

Supplementary Tables

PointByPointResponse

Dr. Michael Hiller
Senior Research Group Leader
Max Planck Institute of Molecular Cell Biology and Genetics
Pfotenhauerstr. 108, Dresden, Germany
Email: hiller@mpi-cbg.de
Phone: +49 351 210-2781
https://www.mpi-cbg.de/hiller

September 10th, 2019

Dear Dr. Edmunds

Thank you very much for considering a revised version of our manuscript “RepeatFiller newly identifies megabases of aligning repetitive sequences and improves annotations of conserved non-exonic elements”.

We would like to thank both reviewers for their constructive comments. We have now addressed all points raised and revised the manuscript. In particular, we have

- generated removed many dependencies from the code base, facilitating the installation by users,
- applied RepeatFiller to fragmented mammalian genomes, which showed that the tool also finds >30 Mb of previously-undetected alignments for fragmented genomes,
- applied RepeatFiller to alignments of birds, reptiles and insects, which showed that RepeatFiller finds less new alignments, but for birds and reptiles still a considerable amount (>1 Mb), suggesting that RepeatFiller can be applied to a wide range of species,
- and investigated which factors influence the amount of added alignments.

New figures or figure panels are: Figure 2B, 3 and Supplementary Figure 1. Supplementary Table 2 is new and new data has been added to Supplementary Table 1. Text changes are highlighted in red font. Please find our point-by-point response to the comments raised by the reviewers uploaded as a separate pdf document.

We hope that our revised manuscript is now acceptable for publication in *GigaScience*. We look forward to hearing from you.

Sincerely,

Michael Hiller