GPAC—Genome Presence/Absence Compiler: A Web Application to Comparatively Visualize Multiple Genome-Level Changes

Angela Noll,*[,1] Norbert Grundmann,[2] Gennady Churakov,[1] Jürgen Brosius,[1] Wojciech Makałowski,[2] and Jürgen Schmitz*[,1]

[1]Institute of Experimental Pathology, ZMBE, University of Münster, Münster, Germany

[2]Institute of Bioinformatics, Faculty of Medicine, University of Münster, Münster, Germany

***Corresponding author:** E-mail: a.noll@uni-muenster.de; jueschm@uni-muenster.de.

Associate editor: Tal Pupko

Abstract

Our understanding of genome-wide and comparative sequence information has been broadened considerably by the databases available from the University of California Santa Cruz (UCSC) Genome Bioinformatics Department. In particular, the identification and visualization of genomic sequences, present in some species but absent in others, led to fundamental insights into gene and genome evolution. However, the UCSC tools currently enable one to visualize orthologous genomic loci for a range of species in only a single locus. For large-scale comparative analyses of such presence/absence patterns a multilocus view would be more desirable. Such a tool would enable us to compare thousands of relevant loci simultaneously and to resolve many different questions about, for example, phylogeny, specific aspects of genome and gene evolution, such as the gain or loss of exons and introns, the emergence of novel transposed elements, nonprotein-coding RNAs, and viral genomic particles. Here, we present the first tool to facilitate the parallel analysis of thousands of genomic loci for cross-species presence/absence patterns based on multiway genome alignments. This genome presence/absence compiler uses annotated or other compilations of coordinates of genomic locations and compiles all presence/absence patterns in a flexible, color-coded table linked to the individual UCSC Genome Browser alignments. We provide examples of the versatile information content of such a screening system especially for 7SL-derived transposed elements, nuclear mitochondrial DNA, DNA transposons, and miRNAs in primates (http://www.bioinformatics.uni-muenster.de/tools/gpac, last accessed October 1, 2014).

Key words: **GPAC, presence/absence visualization, multilocus genome comparison, UCSC Genome Browser, retroposons, exons, introns, numts.**

Introduction

Genomes are the inheritable units passed on from generation to generation since the beginning of life. From this very first moment they continuously accumulated changes and so retained their evolvability in shifting environments over hundreds of million years. Such changes are simple nucleotide mutations, complex genomic rearrangements, duplications, and various random insertions and deletions (indels). Understanding the consequences of all these genomic changes is still a challenge. Now, in the postgenomic era, we start to comprehend the high plasticity of genomes as well as the consequences for organisms mainly by comparative views of the growing quantity of available genomes. The enormous contribution of genomic indels is well illustrated by the high load of transposed elements responsible for about 69% of the human 3-billion letter genomic code (de Koning et al. 2011). To visualize thousands of such genomic changes in one step, we developed the GPAC (genome presence/absence compiler) that screens genome-wide for the presence or absence of insertions among complete alignments of mammalian genomes or those of other species. Such presence/absence patterns (a clear presence in one species and its corresponding absence in an orthologous locus of another species) can be used not only for deriving relationships via shared genomic changes but also for investigating subsequent functional patterns that are correlated to insertion processes.

The National Center for Biotechnology Information (NCBI) is one of the most comprehensive sources for molecular genomic data, with a rigorous coordinate system to structure and compare genomic sequence information. The basic local alignment search tool (BLAST) was developed at NCBI to compare query sequences in a pairwise manner against libraries of subject sequences and to quickly find similar sequences. Founded during the first genome projects and supported by the genomic information of the NCBI database, the University of California, Santa Cruz (UCSC) Bioinformatics Group associated with the Web Miller lab at the Penn State University Center for Comparative Genomics and Bioinformatics, established multispecies sequence comparisons for a broad range of vertebrate and invertebrate taxa by developing

Genome Browser tools to assign annotation tracks (e.g., retrotransposed elements, numts, DNA transposons, miRNAs, exons, and introns). The UCSC BLAST like alignment tool is assigned for species-specific searches, facilitating a fast and accurate search engine with results fully linked to genome loci of various other species and depicted as multiway alignments. However, the UCSC tools currently enable one to visualize orthologous genomic loci for a range of species but only in a single locus mode. In molecular biology, however, it is absolutely necessary to comparatively visualize signals from many loci simultaneously so as to obtain a relevant set of signals for downstream analyses. One criterion for this filtering is the presence of a genomic sequence in a set of species and a clear absence in others. Such presence/absence patterns, for example of retroposed elements, are a valuable retrophylogenomic marker system of the evolutionary relatedness of species (e.g., Kriegs et al. 2006). They can also help to understand the origin and acquisition of novel transcriptomic components (e.g., Krull et al. 2007), resolving novel processes of expression of snoRNA/retroposon hybrids (Schmitz et al. 2008), or to reconstruct the first appearance of a virus in a vertebrate genome (Suh et al. 2013). Here, we present the first GPAC tool to derive comprehensive genome- and species-wide insertion patterns from any list of specific sequence coordinates of the reference species in the multiway genome alignments. The output of GPAC is a fully sortable table that summarizes all presence/absence patterns and link to background information of all individual cases and sequence regions. An automated tree reconstruction indicating the number of informative insertions for each node is available for unambiguous perfect presence/absence patterns (patterns in which there are only single presence ("+") or absence ("−") symbols and no contradictions within a group). After describing the GPAC, we demonstrate its applicability and possible subsequent analysis steps derived from genome-wide investigations in some selected examples of multitaxon comparisons.

Results

GPAC takes advantage of the block structure of the multiple alignment format (MAF) and the rigorous coordinate system of the UCSC database. Currently, GPAC hosts seven multiway alignments (human 46-way, human 100-way, gorilla 11-way, mouse 60-way, opossum 9-way, fruit fly 15-way, and *Caenorhabditis elegans* 6-way) and can be expanded as requested.

Multiple alignments such as the 46-way alignment between human as the leading sequence and 45 other vertebrate genomes can be screened for the presence/absence patterns of a set of selected target coordinates in all 46 species. The MAF alignments, which can be selected from a GPAC pull-down menu (e.g., the human 46-way alignment, about 250 GB and the mouse 60-way alignment, about 240 GB), were downloaded to our server, processed, and indexed to provide a fast, user-friendly GPAC analysis. A local version of GPAC for different computer systems is available on request.

The applicability of GPAC was demonstrated for the following four examples:

Insertion Activities of 7SL-Derived Retroposed Elements in Primates

GPAC is perfectly suited to resolve the origin and evolutionary distribution of transposed elements. This information can be used, for example, to reconstruct a reliable phylogenetic tree of species and to trace the insertion activity patterns of specific element families or subfamilies (Kriegs et al. 2007). Such data were also used to establish a test case for the GPAC application (http://www.bioinformatics.uni-muenster.de/tools/gpac, last accessed October 1, 2014) (figs. 1 and 2*A*). A total of 1,194,734 human 7SL-derived, *Alu*-related elements were located in the UCSC Table Browser (group: Repeats; track: RepeatMasker; filter repName: fossil Alu monomers (FAM)/free left Alu monomers (FLAM)*/free right Alu monomers (FRAM)/Alu*), from which we selected only perfect (defined as all species being assigned to clear presence or absence patterns with just one historical insertion event) or nearly perfect presence/absence patterns from representative subfamilies. For these, we sorted the loci ("Manual" sorting) according to the age of the different 7SL-derived subfamilies starting with FAM, FLAM (FLAM subtypes A, C), FRAM, dimeric Alu Jo, Jb, Sx (subtypes 1, 3), Sq (subtype 2), Sg (subtype 4), Sc (subtype 8), Sp, and Y (subtypes f5, a5) (Quentin 1992; Kapitonov and Jurka 1996). Running 40 selected loci (supplementary table S1, Supplementary Material online) can be sorted according to different criteria for both the investigated loci and the selected species (figs. 1 and 2*A*). Please note the few cases where no perfect pattern (+ or −) could be derived. These cases are described in figure 2*B*.

The 40 selected loci were then used to completely resolve the phylogenetic tree of primates for the available species of the 46-way alignment and to derive by hand a presence/absence tree (fig. 3*A*). GPAC also provides the possibility to directly generate such a marker tree for perfect presence/absence patterns (excluding the cases without clear [+] or [−] signals; supplementary fig. S1, Supplementary Material online). It should be noted that the search for informative markers to resolve this phylogenetic tree started with elements that are present in human (leading sequence of the 46-way alignment); therefore, only markers on the lineage leading to human were detectable.

Appearance of Mitochondrial Inserts in Nuclear Genomes

Following organellar apoptosis, double-stranded fragments of mitochondrial DNA can randomly integrate into the nuclear genome (nuclear mitochondrial DNAs or numts) (Lopez et al. 1994; Hazkani-Covo and Covo 2008). To derive the phylogenetic distribution patterns of primate numts, we searched for clear presence/absence cases by selecting the coordinates of the 766 known human numts compiled by UCSC (Table Browser, group: All Tracks; tracks: NumtS Sequences) and searching for the detailed insertion time in primate evolution. Therefore, we screened all cases for phylogenetically

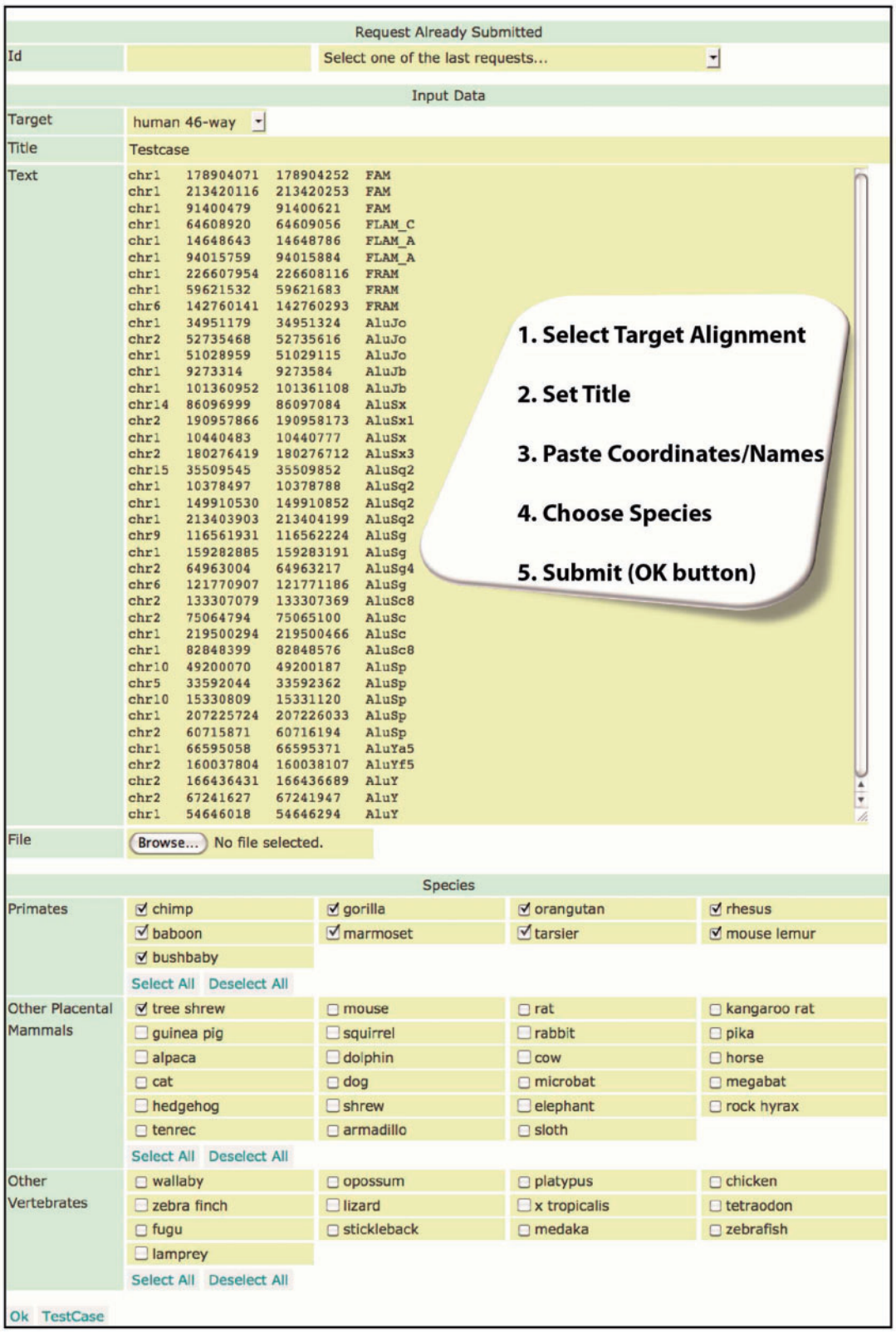
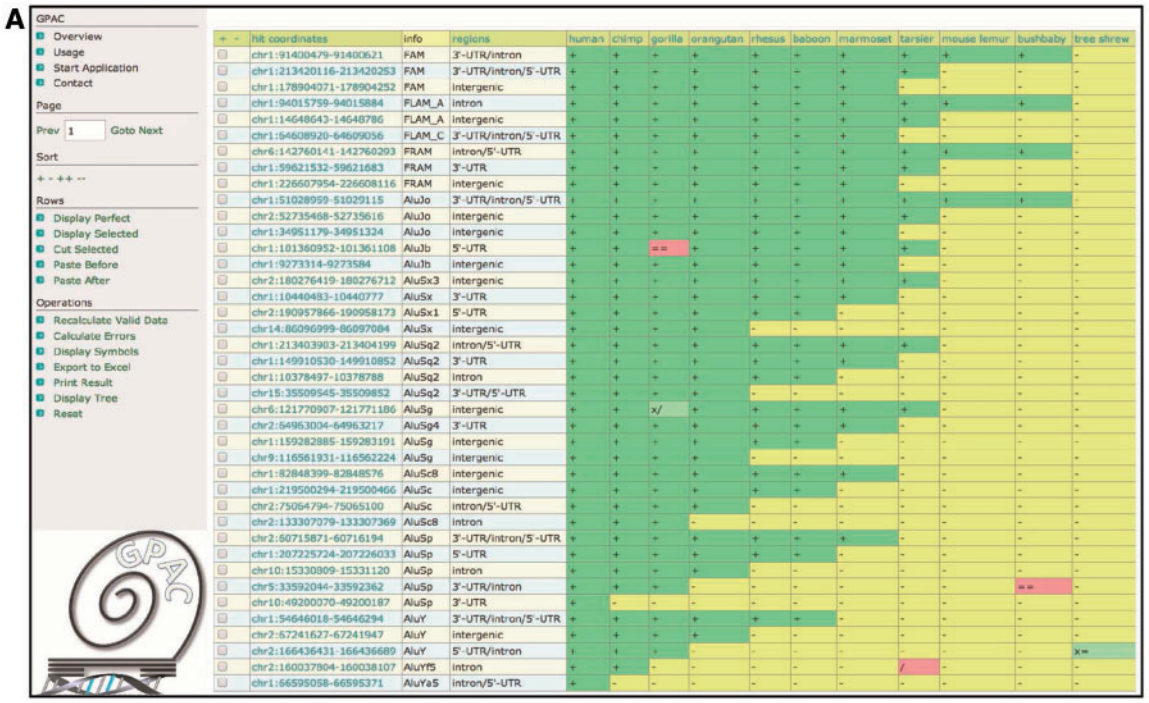


Fig. 1. Entrance screen of the GPAC application written in UCSC notation. The first step of a GPAC analysis is to select the target genome alignment (e.g., the human 46-way MAF sequences). Second, a title is chosen for later recognition. The next entry requires a list of coordinates starting from the chromosomal location and genomic start and end coordinates optionally followed by an insert name. The various species to be examined are then selected. The input is completed and the "OK" button starts the analysis. To demonstrate the functionality of GPAC, we prepared a test case ("TestCase") of 7SL-derived Alu retroposon insertions in primates with the tree shrew as outgroup. Note, for this illustration we selected from hundreds of thousands of insertions only those coordinates containing perfect or nearly perfect insertion patterns (i.e., only unambiguous presence or absence character states) determined during previous GPAC analyses.



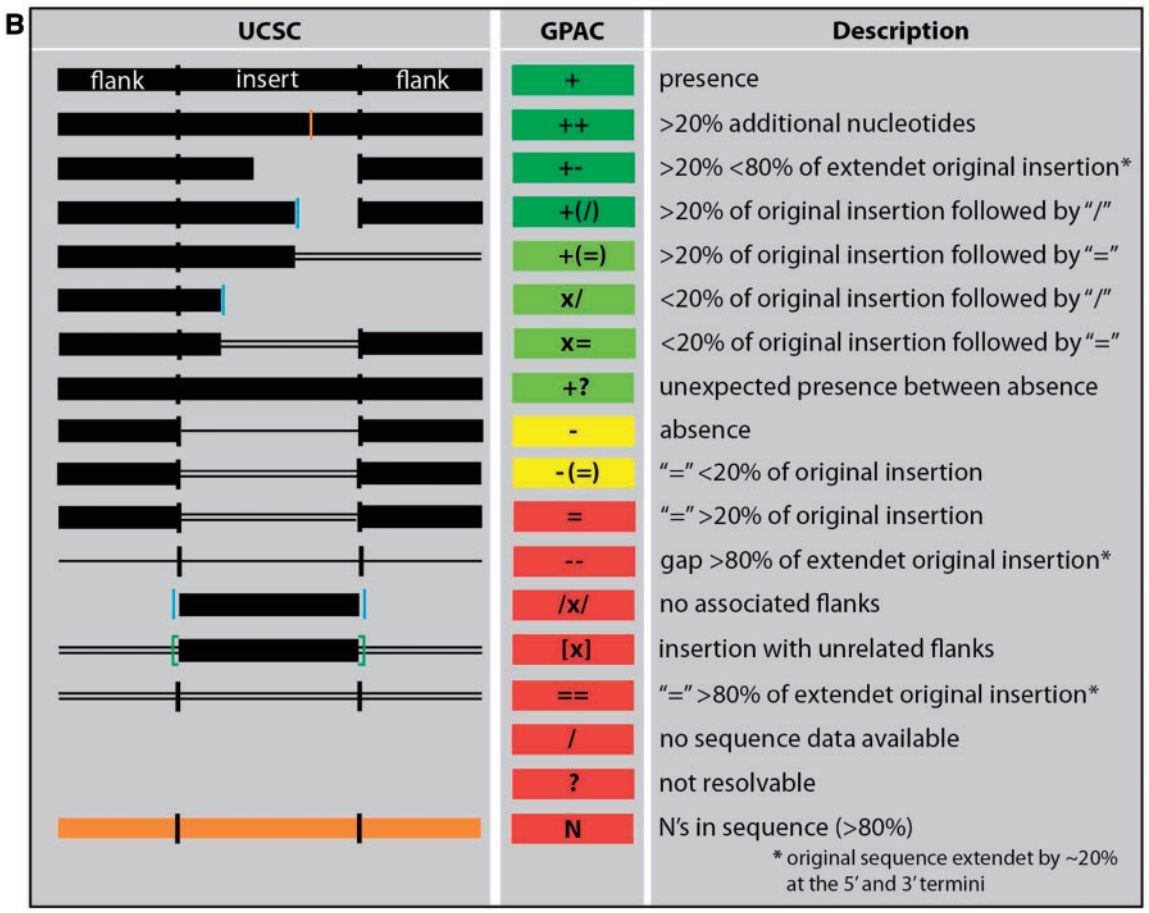


Fig. 2. Gap evaluation and presence/absence symbols. (*A*) Illustrates the usual GPAC output for a test case of nearly perfect presence/absence patterns of 7SL-derived Alu retroposon insertions. The first column serves to select specific elements for sorting. The second and third columns list the coordinates and the user-defined target names. The fourth column defines the genomic region (intronic, intergenic, UTR, etc.) of the considered element. The subsequent columns represent the presence/absence patterns in the selected species. The 40 presence/absence cases for 7SL-derived retroposons shown in the table are sorted according to the age of the elements (in order FAM, FLAM, FRAM, Alu Jo, Alu Jb, Alu Sx, Alu Sq, Alu Sg, Alu Sc, Alu Sp, and Alu Y). The tree shrew was chosen as outgroup. Sorting can also be carried out according to the range of species and coordinates, or according to the number of presence (+) or absence (−) states. Also expanded presence (++) or absence (−−) cases can be selected for deeper analysis.

(continued)

informative presence/absence patterns (54 cases, supplementary table S2, Supplementary Material online).

The GPAC presence/absence pattern of numts shows that, for example, such nuclear insertions accumulated continuously over the 63 My of primate evolution (fig. 3*B*; for dating, Goodman [1999]). The highest number of numts was detected at the terminal branch of human (18), indicating high recent activity of mitochondrial insertions and/or a low divergence of the numts from their original mitochondrial DNAs. Although short internal branches, such as for haplorhines or human plus chimpanzee, accumulated fewer insertions (in both cases just one integration). Remarkably, from the 54 selected cases with clear presence/absence patterns (supplementary table S2, Supplementary Material online), ten show short target site duplications of 3–7 nt. The lengths of the numts vary from 38 (chr3:25508995–25509033) to 2,685 nt (chr2:202077019–202079704). Interestingly, one location contains an insertion of three concatenated numts from different mitochondrial regions (chr3:68708101–68708438; human mtDNA J01415.2 positions: 16086–16187, 4350–4426, 3216–3058; bold in supplementary table S2, Supplementary Material online). In a subsequent analysis of GPAC results from numts and 939 annotated miRNAs provided by UCSC, we cross-compared the genomic coordinates and found two of the numts completely or partially overlapped the miRNA precursor sequences (supplementary table S3, Supplementary Material online). Furthermore, a newly available set of recently discovered miRNAs (2,469 cases; [Friedländer et al. 2014]) was used to search for numt overlaps, and an additional three cases were detected (supplementary table S3, Supplementary Material online). To search for overlaps, we used the coordinates of pre-miRNAs, and an overlap was only counted if at least 20 nt of the mature miRNA was part of it. To our knowledge, our GPAC screening and subsequent coordinate analyses yielded the first report of putative functional miRNAs generated from numts.

Mariner-Derived Element 1 DNA Transposition and Overlapping miRNA Evolution

As an example of the effectiveness of GPAC for examining DNA transposons, we looked at the insertion patterns of Mariner-derived element 1 (MADE1) elements, which are Mariner/Tc1 like nonautonomous DNA transposons found in many species including human. We were interested to see whether they were active over the entire primate history or just at specific internal branches of the primate tree. For this screen, we extracted a total of 7,823 MADE1 sequence coordinates from the UCSC Table Browser (group: Repeats; track: RepeatMasker; filter repName: MADE1), which we used to perform a sample GPAC run for DNA transposons in primates with human as leading species and the tree shrew as outgroup (46-way alignment). The analysis revealed that these elements were highly active in the common ancestor of anthropoids, fitting well to previous results (fig. 3*C*) (Pace and Feschotte 2007).

In a separate GPAC screen (data not presented here), we found that the major wave of miRNA insertions was also in the lineage leading to anthropoids. For this analysis, we screened 939 annotated miRNAs provided by UCSC plus 2,469 newly detected miRNAs from Friedländer et al. (2014). That the majority of both the MADE1 and miRNA insertions were in the common ancestor of anthropoids, inspired us to check for overlaps between these two. Checking the GPAC results for overlapping coordinates (see section below) revealed 282 cases of miRNAs that were derived from MADE1 DNA transposons (fig. 3*C*; supplementary table S4, Supplementary Material online). Such a comparison was motivated due only to the multilocus nature of the results of the GPAC screens.

MADE1-like Tc1-mariner DNA transposons were first described in *C. elegans* (Emmons et al. 1983). Such nonautonomous MADE1 elements are ideally suited to evolve into novel miRNAs. They lack the protein-coding transposase gene but maintain the nearly perfect palindromic terminal inverted repeats. The natural function of the terminal inverted repeats

Fig. 2. Continued

Sequences or groups of sequences can be labeled and cut and pasted to other places. Interesting cases can be marked for a selective view and then reset again to the full set. The graphical view can be exported to an Excel file. Furthermore, the user has the possibility to retrieve a marker tree ("Display Tree") of all perfect presence/absence patterns available in the provided table. All coordinates are linked to the corresponding alignments at UCSC. (*B*) The left column, labeled UCSC, represents the different alignment situations from the UCSC Genome Browser. The second column gives the GPAC symbols for the corresponding presence/absence patterns. The various presence/absence patterns derived from the UCSC MAF-alignments include several differentiated qualities that GPAC filters and converts into simple symbolic patterns. (+): clear presence, (++): the insertion region contains more than 20% additional nucleotides, (+−): only a partial insertion of more than 20% but less than 80% of the extended insertion (extended by flanking 20% additional nucleotides), (+(/)): a partial insertion is truncated after 20% of the original insertion length, (x/): a partial insertion is truncated after <20% of the original insertion length, (+(=)): denotes nonaligning bases after 20% of the original insertion sequence, (x=) nonaligning bases after <20% of the original insertion sequence, (+?): a clear presence in a species while at least two previously presented species in the table show a clear absence, (−): clear absence, (−(=)): the length of the nonalignable sequence (shown by = in the UCSC browser) is <20% of the expected size, (=): the length of the nonalignable sequence (shown by = in the UCSC browser) is > 20% of the expected size, (−−): an expanded gap of > 80% of the extended original insertion size, (= =): nonalignable sequence is > 80% of the expected size, (/X/): the insertion sequence is only present without flanking sequences, [X]: the insert is flanked by locus-unrelated sequences, (/): no sequence data are available, (?): the complete locus cannot be reconstructed, (N): undefined sequences. The last column presents a description of threshold settings for insert/gap allocations. Dark green denotes presence, light green unclear presence, yellow absence, and red uncertain status of insertions.

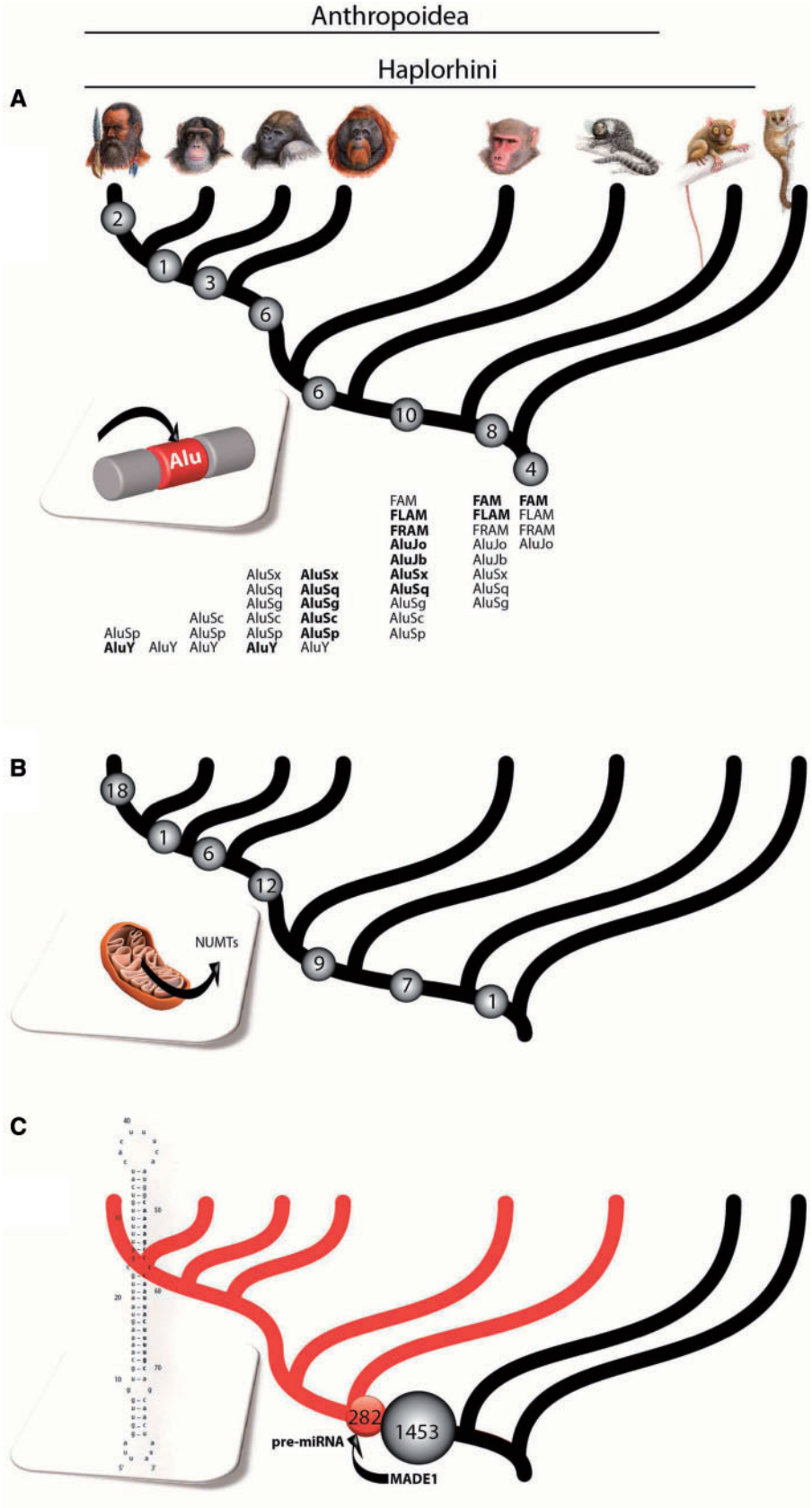


FIG. 3. Phylogenetic representation of GPAC results. Hand reconstructed phylogenetic trees (optional GPAC output) of primates showing the insertions of the major subfamilies of 7SL-derived retroposed elements (*A*), nuclear mitochondrial DNA (*B*), and MADE1 DNA transposons (*C*) were derived from the GPAC presence/absence patterns. (*A*) The insertion activity ranges of different subfamilies of 7SL-derived retroposed elements are not restricted to specific primate branches, but they do mostly overlap speciation events (Quentin 1992). The first waves of insertions involved only monomeric elements. FAMs, for example, were active in the common ancestor of all primates, but continued to be active in the common ancestor of haplorhines

(continued)

is to offer binding sites for a transposase provided by associated autonomous Hsmar1 elements (Morgan 1995). When transcribed, the nearly perfect palindromic structures build stable hairpin loops that can easily evolve into large miRNA families, as was shown for hsa-mir-548 (Piriyapongsa and Jordan 2007).

Overlapping Coordinates

In a comprehensive analysis of all the coordinates so far examined in GPAC searches, we also observed overlaps of coordinates between very different example screenings. This tempted us also to look closer for overlapping coordinates of endogenous retrovirus (ERV)- and 7SL-derived elements (FAM, FLAM, FRAM, and Alu) against miRNAs in browser extensible data (BED)-format (UCSC Table Browser; group: All Tracks; track: sno/miRNA) and additional miRNAs from Friedländer et al. (Friedländer et al. 2014). To check for overlaps, we used BEDTools intersect (option: -wo; Quinlan and Hall [2010]). We identified the mentioned overlaps between miRNAs and 282 MADE1-, 5 numts-, furthermore, 66 ERV-, 2 FAM-, 5 FLAM-, 1 FRAM-, and 35 *Alu*-elements (supplementary tables S3–S5, Supplementary Material online). Transposed elements have quite often been described as donors of miRNAs (Borchert et al. 2006; Piriyapongsa et al. 2007). Most interesting is that, as demonstrated above, the less frequent numts have possibly also become donors of miRNAs.

Genome-Wide Exon/Intron Loss/Gain

The GPAC can also be used to screen for newly generated (gain) or deleted (loss) exons and introns. To conduct such a search, we used the coding exon coordinates of human from the UCSC Table Browser (hg19/GRCh37; group: Genes and Gene Predictions, track: RefSeq Genes, table: refGene, output format: BED; coding exons). We removed multiple hits and those with nonstandard chromosome names (e.g., chr1_gl000191_random or chr6_ssto_hap7) from this list, and used the remaining 202,147 coordinates to analyze the presence/absence patterns of exons within placentals and marsupials as the outgroup. Note that precise or nearly precise exon losses are unlikely. Our initial preliminary screen was for sequence gaps representing missing exons and revealed 13 cases of exon losses within the 46-way vertebrate alignment (supplementary table S6, Supplementary Material online). Most of them (nine) represent losses originating in the exon-flanking introns, probably via recombination of low-complexity intronic sequences. The remaining four cases represent partial exon deletions that continue into the adjacent introns. Most of the deleted exons were located closer to the 3′-end of a gene (seven cases) exposed to less functional impact (Weiner et al. 2006), but also affected the first exons (three cases). In most examples, the GPAC pattern revealed that exon loss was species specific (11 cases). In the other two cases, the exons were lost in the common ancestors of lemurs and rodents. This screen also revealed one exon that was acquired in the ancestor of anthropoids, and another at chr6:10928635–10928683, in which a human medium reiteration frequency interspersed repetitive elements 11C (MER11C) long terminal repeat (LTR) element inserted into a splice site of exon 17 of the synaptonemal complex protein 2-like gene (*SYCP2L*), elongating exon 17 by 30 nt. The inserted MER11C element provided an additional splice site and was thereby responsible for a new exon of 48 nt. GPAC indicates a hominoid origin of the insertion and exonization events. Along with the new exon, a new intron also evolved (fig. 4). The GPAC-detected losses of exons resemble natural forms of knockouts.

We also used GPAC for a mammalian-wide inspection of 217,993 intron coordinates from human (hg19/GRCh37) and 196,478 from mouse (mm10/GRCm38), and thereby identified 93 and 64 cases, respectively, with clear presence/absence boundaries for complete intron deletions/insertions, including their splice sites (supplementary tables S7 and S8, Supplementary Material online). This set of data represents an interesting mosaic of intron losses and gains in many mammalian species and species groups. From the 46-way human-based alignment, GPAC revealed, among others, one case of clear intron loss for all rodents, five for Muridae (mouse- and rat-related rodent species), one for marsupials, two for lagomorphs, and one intron gain for hominoids (supplementary table S7, Supplementary Material online). From the 91 total intron losses and 2 gains discovered from the screen of the 46-way human-based alignment, 16 were the last, and 4 were the first introns of a gene. All others were located between the first and last exons or were part of the 5′- or 3′-UTR. Most intron losses were species-specific (e.g., 4 for marmoset, 5 for bush baby, 6 for tree shrew, 6 for mouse, and 13 for rat; see supplementary table S7, Supplementary Material online). From the 60-way alignment based on the mouse genome, we found 64 intron losses most of them species-specific for rat (33). Among others, there were five intron losses in hystricognathi (naked mole rat and guinea

Fig. 3. Continued

and anthropoids. The youngest and still active dimeric elements are Alu Y with a broad range of activity from the common ancestor of Old World monkey to human-specific events. The predominant element insertions, or most actively inserted elements are indicated by bold letters. Manually inserted below the tree are the names of the inserted elements corresponding to the circles at the various branch points of the tree (not provided by GPAC). The numbers in the circles refer to the number of different elements with unambiguous insertion patterns during the given time. (*B*) The nuclear copies of mitochondrial DNA (numts) that we examined inserted continuously over the entire evolution of primate species. Numbers represent the numbers of insertions with clear presence/absence representations. (*C*) MADE1 DNA transposons were broadly active on the lineage leading to anthropoid primates. The structural sequence represents the hsa-mir-54 pre-miRNA. The red part of the tree indicates the activity-range of the MADE1-derived miRNAs.

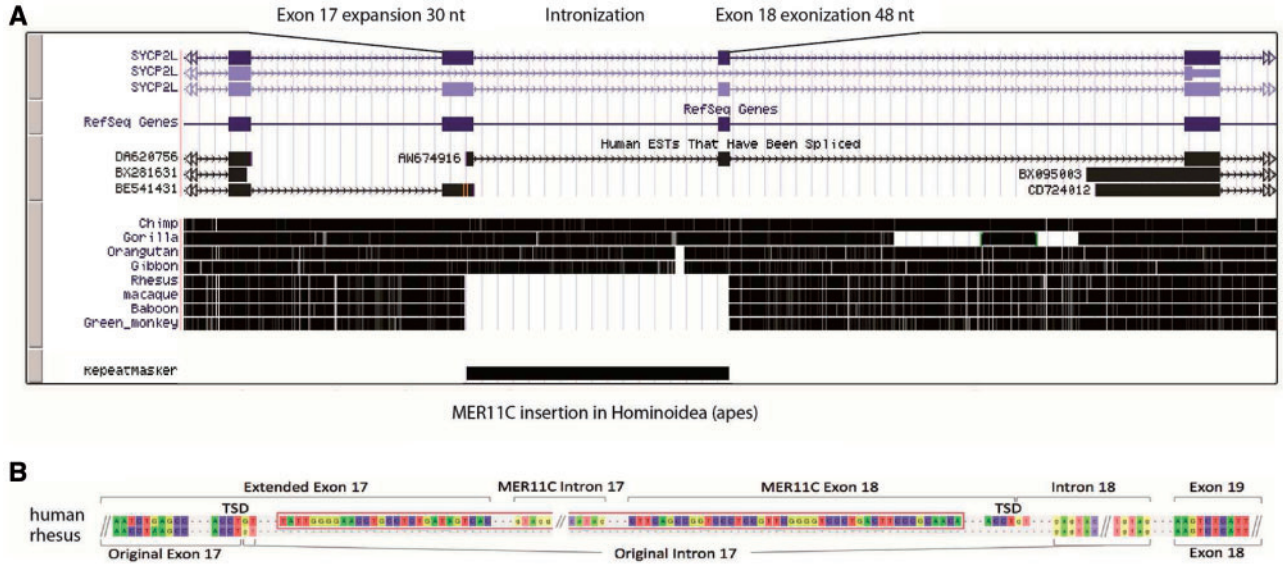


FIG. 4. Exon/intron gain identified from a GPAC pattern. (*A*) The UCSC Genome Browser view presents a section of the human *SYCP2L* gene (human coordinates: chr6:10,926,575–10,930,748). Following the insertion of an MER11C retroposon, an expanded exon 17 (additional 30 nt) was formed, and a new intron and a new exon 18 (48 nt) arose from the retroposed element. The upper part of the Genome Browser screen shows the gene structure in the sense orientation for the RefSeq Genes, aligned EST reports, and selected species from the vertebrate Multiz Alignment & Conservation (100 species) pattern. The RepeatMasker-determined MER11C insertion is located at the bottom of the conservation pattern. (*B*) The sequences spanning exon 17 and exon 19 of *SYCP2L* in human (top) and the orthologous sequences for rhesus (bottom). Note that the sequence in rhesus represents the gene structure before insertion of the MER11C retroposon. Therefore, exon 17 is truncated and the newly exonized exon 18 in human does not exist in rhesus. The MER11C element is boxed and flanked by short target site duplications (TSD) characteristic of LTR-derived elements. The original GT-splice site is now part of the coding exon 17 in human. The new exon 18 in human acquired the 3′-splice site from the flanking intronic sequence (intron 18) and is part of the right TSD. The human open reading frame of the gene is still intact after acquisition of the 78 new coding nucleotides. Intronic sequences are shown in lower case letters. The two internal slashes denote 1,026 and 1,885 nt not shown. EST, expressed sequence tag.

pig) and two within lagomorphs (supplementary table S8, Supplementary Material online).

In *Drosophila*, Coulombe-Huntington and Majewski (2007) described an elaborate biocomputational strategy by which they recovered 213 intron gains and 1,754 losses, numbers which have so far been found only in *Drosophila* species. With GPAC, such an analysis is easy, quick, extremely reliable, and performed without any bioinformatics experience. In a test run of GPAC, we retrieved most previously described examples of trustworthy mobilizations of complete intronic regions in 12 *Drosophila* species and additional cases (491/1,620 cases of gains/losses; supplementary table S9, Supplementary Material online). Interestingly, a comparison of intron loss in *Drosophila* with that in some outgroup species (mosquito, honey bee, and flour beetle) revealed numerous cases of independent intron losses in mosquito and honey bee (data not shown). Thus, based on our GPAC patterns the postulated mechanisms of intron loss in *Drosophila* (Coulombe-Huntington and Majewski 2007) can possibly be transferred to other insect groups. Using the specified sorting option (−−), GPAC also enabled us to detect an enormous number of imprecise exon losses/gains (about 100 cases; supplementary table S10, Supplementary Material online) as well as gene loss and gain during the evolution of *Drosophila* species (about 30 cases; supplementary table S10, Supplementary Material online).

Discussion

GPAC is a novel, easy-to-use application for high-throughput screening of presence/absence insertion patterns of genomic elements. The tool requires nucleotide coordinates of a reference species to search for orthologous regions in a preset database of query species based on the UCSC multigenome alignments. The user receives a graphical representation of the presence/absence patterns in table format with direct links to the corresponding individual sequence views in the UCSC Genome Browser. Furthermore, GPAC provides the possibility to receive the phylogenetic distribution of perfect presence/absence loci in a tree topology. The user-friendly web application provides information about the presence/absence patterns of acquired functional or nonfunctional components of the genome. Such knowledge can be used, for example, for phylogenomic (e.g., retrophylogenomic) analyses, elucidation of the insertion activities of specific DNA or RNA transposons, and the evolution of novel, inserted functional components, such as exons, introns, and miRNAs.

As demonstrated for 7SL-derived elements, GPAC large-scale presence/absence analyses can be carried out to characterize the insertion activities of ancient and recent transposons. We screened for a minor fraction (40 selected cases) of *Alu*-like sequences with nearly perfect presence/absence configurations for all primates from the 46-way,

human-based genome alignment and a tree shrew outgroup. In so doing, we derived a pattern of the full evolutionary spread of these elements (see also Kapitonov and Jurka 1996) as well as a completely resolved divergence pattern of the branch leading to human (fig. 3A). These data are in full agreement with established phylogenetic patterns, for example, those based on large-scale sequence analyses (Churakov et al. 2010; Perelman et al. 2011). Numts are another frequent form of eukaryotic nuclear insertions, most of which usually are transcriptionally inactive (dead-on-arrival) or at the most cotranscribed following insertion. However, the processes of their transfer and insertion are mostly unknown (Tsuji et al. 2012). GPAC enables the gathering of large amounts of data necessary for understanding the phylogenetic distributions of such insertions. By comparing the flanking sequences of numts, we suggest two possible ways that numts might be inserted: 1) via nonhomologous end joining repair (Hazkani-Covo and Covo 2008) and 2) hitherto unknown, possibly via retroposition, that leaves behind short flanking direct repeats (Hindmarsh and Leis 1999). Furthermore, by subsequent comparison of chromosomal coordinates, we found sequence overlaps of five miRNAs with numts, demonstrating that the latter are not necessarily transcriptionally inactive. To our knowledge, this is the first report of numt-derived candidate miRNAs. The GPAC patterns also enabled us to show that numts have accumulated continuously over the entire evolutionary divergence time of primates, which correlates well with the length of their internal and terminal branches.

GPAC provides an opportunity to collect high-throughput data that can be used to elucidate mechanistic processes of the evolution of novel genomic modules (for the modular architecture of genes see also Brosius [2009]). This we demonstrated for the insertion of MADE1 DNA transposons that are exapted into experimentally verified human miRNA sequences. The nonautonomous MADE1s are well known for their insertions predominantly close to genes (Tu 1997) and their perfect palindromic structures are similar to pre-miRNAs, and therefore, by nature, ideally located and suited to form active miRNAs. Over the last 63 My of primate evolution, they formed important active miRNA units for posttranscriptional regulation (Piriyapongsa and Jordan 2007). Such elements were probably transformed in a process of self-regulation to reduce their own transposon activity. Interestingly, MADE1 elements are significantly less frequently inserted into the 3′-UTR than into the 5′-UTR sequences of their human host genes (27 of 7,794 MADE1 elements in 3′ and 281 in 5′-UTRs), indicating strong selection against their presence in the 3′-UTR. Furthermore, their potential to form 3′-UTR miRNA target sites might inhibit the expression of the host gene. With GPAC screens and subsequent analyses of overlapping genomic coordinates, we detected 282 miRNAs derived from MADE1, most of them undetected in the more than 2,000 novel human miRNAs recently described (Friedländer et al. 2014). MADE1 elements have a relatively recent and narrow time span of insertion activity in the common ancestor of anthropoids. For hsa-mir-548, more than 3,500 potential target genes, some of them involved in cancer regulation, have been described; also the *Alu*-derived miRNA hsa-mir-566 shows 1,184 predicted target sites that derived mostly from *Alu*-element insertions (Piriyapongsa et al. 2007).

The gains of exons and introns in the lineages leading to human or mouse are easily traceable with GPAC and can reveal interesting evolutionary processes. This was demonstrated for the MER11C element insertion into an existing splice-site in human, which destroyed the site and led to an expansion of the corresponding exon and the generation of an addition exon together with a new LTR-derived intron (fig. 4). Even if introns are under less selection pressure than exons (Smith and Hurst 1998)—except for those containing highly conserved regulatory regions (Duret and Bucher 1997)—it is possible, using GPAC, to obtain clear intronic presence/absence patterns. The gain or loss of introns and exons still represents a mysterious chapter in genome evolution (Jeffares et al. 2006), and a systematic analysis of GPAC data from hundreds of thousands of corresponding intron/exon coordinates might produce the long-awaited insights needed to unravel this mystery.

Conclusions

GPAC is an easy, fast, flexible, and robust application for visualizing genome-level changes by a simple presence (+) absence (−) code in a multispecies and multilocus environment perfectly suited for genome-level analyses.

A current limitation of GPAC is the restricted set of MAF multiway alignments available. In the GPAC version submitted here, we incorporated both the 46-way and the new 100-way alignments with human as the leading sequence. As a further hominoid sample, the 11-way alignment with gorilla as the leading species is provided. For rodents, we provide access to the analysis of the mouse 60-way alignment and for marsupials the only available 9-way alignment with opossum. The multiway alignments of *C. elegans* and *Drosophila* are also incorporated into the current GPAC version. We are in the process of engineering an automated multiple two-way genome aligner to connect this tool with a next-generation GPAC for an unrestricted compilation and analysis of any genome data. An initial strategy of using two-way alignments to gain multilevel presence/absence information is outlined in Hartig et al. (2013).

The GPAC tool maintains the UCSC notation so that it can be easily incorporated into the UCSC Genome Browser to fill an important analysis gap and is only restricted by the available number and quality of annotated sequence data and MAF alignments. With the forthcoming vast quantity of genome data from the 10K vertebrate genome project (https://genome10k.soe.ucsc.edu, last accessed October 1, 2014), GPAC will establish a new dimension for gaining quick and easy access to comparative, multigenome information, essential in the postgenomic era of high-throughput data analyses.

Materials and Methods

The GPAC is designed to comparatively screen multiple genome regions for the presence or absence of specific sequences among a preselected number of species. GPAC

requires a list of corresponding genomic coordinates for such specific sequences (e.g., retroposed elements, introns, exons, miRNAs). It then extracts information about the presence or absence of such components from available multiway alignments (http://genome.ucsc.edu/FAQ/FAQformat.html#format5, last accessed October 1, 2014) and transforms this information into simple symbols stored in a sortable, color-coded table that is directly linked to corresponding UCSC genome alignments. The various possibilities of clear (+ and −) and unclear insertion states range from additional sequences in presence regions (e.g., ++), missing parts or complete flanking regions (e.g., +(=)), to no available sequence data (/) (for a complete interpretation of different insertion scenarios see the legend to figure 2*B* and the MAF-format description (http://genome.ucsc.edu/FAQ/FAQformat.html#format5, last accessed October 1, 2014). A more detailed case study is provided in supplementary fig. S3, Supplementary Material online.

The core of the GPAC algorithm, the Calculator, is written in the Python programming language (version 2.7). To simplify the handling of this program, it is encapsulated in a web environment using Perl and JavaScript to exchange and visualize data from the core part. A Perl script, the Generator, is responsible for receiving, checking, and storing request data and creating web pages. When the user opens the first input page and selects an already calculated request by entering the identification number (ID) of a stored run, the results table is immediately displayed. To avoid unnecessary processing by the server, a JavaScript program, the Presenter, sorts, displays, and exports the data directly in a web browser at any local computer. The calculation, generation, and presentation steps are executed on different machines, which makes the processes very flexible and relieves the server.

The first step in applying GPAC is to call up the webpage (http://www.bioinformatics.uni-muenster.de/tools/gpac, last accessed October 1, 2014; fig. 1) to enter input data. After entering the data, clicking on "OK" sends it to a server, queued for calculation in order of arrival. Every request gets an individual ID, so one can retrieve previously sent requests easily by providing the ID. It is also possible to name the runs so it is easier to refer the results to an input data set. The titles and IDs of the last ten runs are kept in the browser and are accessible from a selection box. Results are kept on the server for 7 days and then automatically removed. The data can also be exported as an excel file for long-term storage. It should be noted that JavaScript and Cookies must be enabled in the browser for the application to work properly.

To display large data sets the result table is split into fragments of 100 lines. You can navigate through the pages using a "Prev" and "Next" button or go directly to any entered page number.

Input

Subject

The current preset subject database for searching for presence/absence patterns in GPAC is a 46-species assembly to the human genome hg19/GRCh37 (February 2009, multiz46way alignment; http://hgdownload.soe.ucsc.edu/goldenPath/hg19/multiz46way/maf, last accessed October 1, 2014) in MAF. Alternatively, one can choose an extended 100-way alignment based on human hg19, (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/multiz100way/maf, last accessed October 1, 2014) and an 11-way alignment led by *Gorilla gorilla* (gorGor2, May 2011; http://hgdownload.soe.ucsc.edu/goldenPath/gorGor3/multiz11way/). A 60-way alignment based on the mouse genome mm10/GRCm38 (December 2011, multiz60way alignment; http://hgdownload.soe.ucsc.edu/goldenPath/mm10/multiz60way/maf, last accessed October 1, 2014) is available to visualize presence/absence patterns on the lineage leading to the mouse, and a nine-way alignment of *Monodelphis domestica* (October 2006, multiz9way alignment; monDom5; http://hgdownload.soe.ucsc.edu/goldenPath/monDom5/multiz9way/maf, last accessed October 1, 2014) for marsupials. Invertebrates are represented by a seven-way alignment of *C. elegans* (May 2008, multiz6way, ce6; http://hgdownload.soe.ucsc.edu/goldenPath/ce6/multiz6way, last accessed October 1, 2014) and a 15-way alignment of *Drosophila melanogaster* (April 2006, multiz15way alignment, dm3; http://hgdownload.soe.ucsc.edu/goldenPath/dm3/multiz15way, last accessed October 1, 2014). This list will be extended and updated regularly.

Query

The query is a user-compiled list of coordinates derived from one of the leader sequences of the multiway alignments (e.g., human, in BED-format for multiz46way; http://genome.ucsc.edu/FAQ/FAQformat.html#format1, last accessed October 1, 2014) with an optional tab-separated query name (e.g., chr1➛12345➛23456➛name). It is also possible to use the format provided by the UCSC Genome Browser (e.g., chr1:12345-23456➛name; http://genome.ucsc.edu, last accessed October 1, 2014). Before starting a run, uncommon chromosome names (e.g., chr1_gl000191_random or chr6_ssto_hap7) that cause warning messages are removed.

The Calculator Program

The Calculator first extracts all MAF-blocks (supplementary fig. S2, Supplementary Material online) associated with the query coordinates. Orthologous sequence regions in all selected species are then evaluated for the presence/absence of the given loci with the help of the sequence status information stored in the extracted MAF-blocks. Within each block, the multiple way alignment of a given genomic region is provided at the DNA level (within the so called "s"-lines). Sequences present before and after this block of the aligned species ("i"-lines) and the lengths of the gaps spanning the block ("e"-lines) are also given (http://genome.ucsc.edu/FAQ/FAQformat.html#format5, last accessed October 1, 2014). Parsing and summarizing this information for an approximately 40% extended query-region (an ~20% upstream and ~20% downstream expansion for additional quality control of the exact query sequence boundary) provides the presence/absence status for each selected species. The results are stored in a color-coded table with symbols reflecting the different

presence/absence patterns (fig. 2). Please see the legend to figure 2*B* for a complete listing and definitions of the various symbol codes.

In addition to the presence/absence status of specific coordinates, it is also possible to determine the genomic structures (e.g., exons, introns, 5′-UTR, 3′-UTR, intergenic region) present at user-defined coordinates. To accomplish this, coding sequence and exon frame information are extracted from the UCSC Table Browser (group: Genes and Gene Predictions; track: Ensembl Genes/RefSeq Genes/UCSC Genes; table: ensGene/refGene/knownGene; output format: All fields from selected table; http://genome.ucsc.edu/cgi-bin/hgTables?command=start, last accessed October 1, 2014) and displayed in the results table under "regions" (fig. 2*A*).

Options for the Result Table

The "Sort" option:

"+" sorts the rows by the maximal number of "+", "++", "+−," and "+(/)".
"−" sorts for the maximal number of "−" and "− (=)" per row.
"−−" and "++" sorts for larger deletions or insertions.

The "Rows" option:

"Display Perfect": Selects those rows of species that show only clear character presence ("+") or absence ("−") states.
"Display Selected": Displays only marked rows.
"Cut Selected," "Paste Before," and "Paste After" provide the possibility to move marked rows to other places in the table or remove them.

The "Operations" option:

"Recalculate Valid Data": Recalculates the current data set with possible new settings (e.g., a new species selection).
"Calculate Errors": Offers the possibility to recalculate missing invalid data.
"Display Symbols": Shows all symbols from figure 2*B*.
"Export to Excel": Exports the values of the displayed table directly into Excel-, Libre- or OpenOffice-files.
"Print Result": Prints the results. To avoid fragmentation of the table a new window is opened.
"Display Tree": Presents a phylogenetic tree of the selected species and the numbers of supporting perfect presence/absence markers. For downloading, the tree is represented in the Newick tree format along with the numbers of supportive markers.
"Reset": Redisplays the original result of the full data set.

A click on any species will move this to the first position in the list of species (fig. 2*A*).

Tree Generation

To display a phylogenetic tree for the selected species and to map the informative presence/absence markers, we generated a phylogenetic tree in Newick format for each multiway-alignment provided. To prune the tree to the user-selected species, we used the phylogenetic model "tree_doctor" from the software package PHAST (Hubisz et al. 2010). The numbers of markers for each node are taken from the GPAC output table. Only genomic loci that harbor 1) only "+" and "−" signals and 2) only one "±" boundary indicating the time of the appearance of the element are incorporated into the tree. The numbers of markers were introduced into the Newick tree replacing the commonly used bootstrap values. For subsequent analyses, the Newick tree plus mapped markers can be downloaded and is visualized in a simple graphical topology (for visualization see Smits and Ouverney 2010; supplementary fig. S1, Supplementary Material online).

Supplementary Material

Supplementary figures S1–S3 and tables S1–S10 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This paper is dedicated to the memory of Jerzy Jurka who combined scientific progress in the field of transposed elements with remarkable fairness and honesty in all scientific interactions. This work was supported by the Deutsche Forschungsgemeinschaft (SCHM1469/3-2), the Medical Faculty of the University of Münster, and the Münster Graduate School of Evolution (MGSE). The authors are very thankful to Marsha Bundman for her careful editing of the manuscript. The primate paintings were provided by Jón Baldur Hlíðberg. The authors disclose no competing financial interests.

References

Borchert GM, Lanier W, Davidson BL. 2006. RNA polymerase III transcribes human microRNAs. *Nat Struct Mol Biol.* 13:1097–1101.
Brosius J. 2009. The fragmented gene. *Ann N Y Acad Sci.* 1178:186–193.
Coulombe-Huntington J, Majewski J. 2007. Intron loss and gain in *Drosophila*. *Mol Biol Evol.* 24:2842–2850.
Churakov G, Grundmann N, Kuritzin A, Brosius J, Makalowski W, Schmitz J. 2010. A novel web-based TinT application and the chronology of the Primate Alu retroposon activity. *BMC Evol Biol.* 10:376.
de Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.* 12:e1002384.
Duret L, Bucher P. 1997. Searching for regulatory elements in human noncoding sequences. *Curr Opin Struct Biol.* 7:399–406.
Emmons SW, Yesner L, Ruan KS, Katzenberg D. 1983. Evidence for a transposon in *Caenorhabditis elegans*. *Cell* 32:55–65.
Friedländer MR, Lizano E, Houben AJS, Bezdan D, Báñez-Coronel M, Kudla G, Mateu-Huertas E, Kagerbauer B, González J, Chen KC, et al. 2014. Evidence for the biogenesis of more than 1,000 novel human microRNAs. *Genome Biol.* 15:R57.
Goodman M. 1999. The genomic record of Humankind's evolutionary roots. *Am J Hum Genet.* 64:31–39.
Hartig G, Churakov G, Warren WC, Brosius J, Makalowski W, Schmitz J. 2013. Retrophylogenomics place tarsiers on the evolutionary branch of anthropoids. *Sci Rep.* 3:1756.
Hazkani-Covo E, Covo S. 2008. Numt-mediated double-strand break repair mitigates deletions during primate genome evolution. *PLoS Genet.* 4:e1000237.

Hindmarsh P, Leis J. 1999. Retroviral DNA integration. *Microbiol Mol Biol Rev.* 63:836–843.
Hubisz MJ, Pollard KS, Siepel A. 2010. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief Bioinform.* 12:41–51.
Jeffares DC, Mourier T, Penny D. 2006. The biology of intron gain and loss. *Trends Genet.* 22:16–22.
Kapitonov V, Jurka J. 1996. The age of Alu subfamilies. *J Mol Evol.* 42: 59–65.
Kriegs JO, Churakov G, Jurka J, Brosius J, Schmitz J. 2007. Evolutionary history of 7SL RNA-derived SINEs in Supraprimates. *Trends Genet.* 23:158–161.
Kriegs JO, Churakov G, Kiefmann M, Jordan U, Brosius J, Schmitz J. 2006. Retroposed elements as archives for the evolutionary history of placental mammals. *PLoS Biol.* 4:e91.
Krull M, Petrusma M, Makalowski W, Brosius J, Schmitz J. 2007. Functional persistence of exonized mammalian-wide interspersed repeat elements (MIRs). *Genome Res.* 17:1139–1145.
Lopez JV, Yuhki N, Masuda R, Modi W, O'Brien SJ. 1994. Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. *J Mol Evol.* 39: 174–190.
Morgan GT. 1995. Identification in the human genome of mobile elements spread by DNA-mediated transposition. *J Mol Evol.* 254: 1–5.
Pace JK 2nd, Feschotte C. 2007. The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. *Genome Res.* 17:422–432.
Perelman P, Johnson WE, Roos C, Seuanez HN, Horvath JE, Moreira MA, Kessing B, Pontius J, Roelke M, Rumpler Y, et al. 2011. A molecular phylogeny of living primates. *PLoS Genet.* 7:e1001342.
Piriyapongsa J, Jordan IK. 2007. A family of human microRNA genes from miniature inverted-repeat transposable elements. *PLoS One* 2: e203.
Piriyapongsa J, Marino-Ramirez L, Jordan IK. 2007. Origin and evolution of human microRNAs from transposable elements. *Genetics* 176: 1323–1337.
Quentin Y. 1992. Origin of the Alu family: a family of Alu-like monomers gave birth to the left and the right arms of the Alu elements. *Nucleic Acids Res.* 20:3397–3401.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842.
Schmitz J, Zemann A, Churakov G, Kuhl H, Grutzner F, Reinhardt R, Brosius J. 2008. Retroposed SNOfall—a mammalian-wide comparison of platypus snoRNAs. *Genome Res.* 18:1005–1010.
Smith NG, Hurst LD. 1998. Sensitivity of patterns of molecular evolution to alterations in methodology: a critique of Hughes and Yeager. *J Mol Evol.* 47:493–500.
Smits SA, Ouverney CC. 2010. jsPhyloSVG: a javascript library for visualizing interactive and vector-based phylogenetic trees on the web. *PLoS One* 5(8):e12267.
Suh A, Brosius J, Schmitz J, Kriegs JO. 2013. The genome of a Mesozoic paleovirus reveals the evolution of hepatitis B viruses. *Nat Commun.* 4:1791.
Tsuji J, Frith MC, Tomii K, Horton P. 2012. Mammalian NUMT insertion is non-random. *Nucleic Acids Res.* 40:9073–9088.
Tu Z. 1997. Three novel families of miniature inverted-repeat transposable elements are associated with genes of the yellow fever mosquito, *Aedes aegypti*. *Proc Natl Acad Sci U S A.* 94:7475–7480.
Weiner J 3rd, Beaussart F, Bornberg-Bauer E. 2006. Domain deletions and substitutions in the modular protein evolution. *FEBS J.* 273:2037–2047.