The Wide Distribution and Horizontal Transfers of Beta Satellite DNA in Eukaryotes

Yabin Guo (guoyb9@sysu.edu.cn)
Sun Yat-Sen University https://orcid.org/0000-0001-8316-8527
Jiawen Yang
Sun Yat-Sen University
Bin Yuan
Hubei Academy of Agricultural Sciences
Yu Wu
Sun Yat-sen University
Meiyu Li
Sun Yat-Sen University
Jian Li
Sun Yat-Sen University
Donglin Xu
Guangzhou Academy of Agricultural Sciences
Zeng-hong Gao
Sun Yat-Sen University
Guangwei Ma
Sun Yat-Sen University
Yiting Zhou
Sun Yat-Sen University
Yachao Zuo
Sun Yat-Sen University
Jin Wang
Sun Yat-Sen University

Research article

Keywords: Beta satellite DNA, Sau3A sequences, Eukaryotes, Horizontal gene transfer, Primates

Posted Date: November 15th, 2019

DOI: https://doi.org/10.21203/rs.2.17287/v1

Abstract

Beta satellite DNA (satDNA) sequences, also known as Sau3A sequences, are repeated DNA elements reported in human and primate genomes. Beta satDNAs may play roles in genome stability and chromosome segregation during mitosis. It is previously thought that beta satDNAs originated in old world monkeys and bursted in great apes. However, global and high-throughput studies on beta satDNAs are still absent.Results: In this study, we searched 7,821 genome assemblies of 3,767 eukaryotic species and found that beta satDNAs actually are widely distributed across eukaryotes. The four major branches of eukaryotes, animals, fungi, plants and Harosa/SAR, all have multiple clades containing beta satDNAs. These results were also confirmed by searching whole genome sequencing data (SRA) and PCR assay. Beta satDNA might have originated during the early evolution of eukaryotes. The widely patchy distribution of beta satDNAs across eukaryotes presents a typical scenario of multiple horizontal transfers (HT). In contrast, beta satDNA sequences were found in all the primate clades, Primatomorpha and Euarchonta, indicating an origin in the common ancestor and vertical transfers thereafter. Besides in eukaryotes, beta satDNAs were even found in some archaea and bacteria, which should have been acquired from eukaryotes via HTs.Conclusion:Beta satDNAs widely exist in eukaryotes. The current distribution landscape of beta satDNA is the result of countless HTs. Our study shows for the first time that satellite DNAs can also undergo HT, and will provide new ideas for the future investigations in HT/HGT field.Keywords: Beta satellite DNA, Sau3A sequences, Eukaryotes, Horizontal gene transfer, Primates

Background

The genomes of eukaryotes comprise large tracts of repeated sequences, including satellite DNAs (satDNAs), minisatellite, microsatellite sequences and transposable elements (TEs) [1]. Highly homogenized arrays of tandem repeats, known as satDNAs, are enriched in centromeric, pericentromeric, subtelomeric regions and interstitial positions [2, 3]. SatDNAs have recently been reconsidered to have various functions, such as playing roles in genome stability, chromosome segregation [4] and even gene regulations [5]. The changes in the copy number of repetitive sequences in genomic DNA are important causes of hereditary disorders [6-8]. The insertion of 18-beta satDNA unit in the gene coding a transmembraine serine protease causes congenital and childhood onset autosomal recessive deafness [9]. The global DNA hypomethylation frequently observed in cancers is mostly taken place at satDNAs [8]. The loss of the BRCA1 tumor suppressor gene provokes satDNA derepression in breast and ovarian tumors in both mice and humans [10]. Facioscapulohumeral muscular dystrophy (FSHD), a autosomal dominant hereditary disease, is associated with the macrosatellite (D4Z4) and beta satellite (4qA allele) DNA sequences [8].

Beta satDNAs were considered to be unique in primates. The basic units of beta satDNAs are 68 bp long with a higher GC content. Beta satDNAs are also known as Sau3A sequences for the presence of a Sau3A restriction site within nearly every single unit [11]. In the human genome, beta satDNAs have been identified at the pericentromeric regions of chromosomes 1, 9, and Y [12], as well as in acrocentric

chromosomes 13, 14, 15, 21 and 22 [13] and in chromosome 19p12 [14]. Beta satDNAs were also described in great apes (human, chimpanzee, gorilla, and orangutan), lesser apes [15] and old world monkeys [16]. However, most of these studies were performed before next generation sequencing was utilized and the amount of sequences studied was quite limited. The proportion of beta satDNA in human genome, the distribution range of beta satDNAs in the tree of life and how they emerged in primates largely remain elusive.

In this study, we searched almost all the genome assemblies, including genome assemblies of eukaryotes, prokaryotes, viruses and organelles, in the NCBI Genome database, and revealed that beta satDNAs are widely distributed across eukaryotes. The distribution landscape of beta satDNAs presents a scenario of multiple horizontal transfers (HT/HGT) during evolution.

Results

Beta satDNA sequences were identified in genomes of multiple eukaryotic taxa

To study the distribution of beta satDNAs in different species, we BLASTed human beta satDNA sequences online against the nucleotide collection of NCBI. In addition to primates that known to harbor beta satDNAs, significant hits were found in *Spirometra erinaceieuropaei*, *Onchocerca flexuosa*, *Enterobius vermicularis*, *Bos mutus* and *Nicotiana tabacum*. *S. erinaceieuropaei, O. flexuosa* and *E. vermicularis* are endoparasites of human and other mammals. The presence of beta satDNA in these species indicates HTs. To our great surprise, hit was also found in a plant, *N. tabacum*. This result suggests that the distribution of beta satDNAs could be much wider than what we have known. It is necessary to perform a global investigation for beta satDNA sequences in eukaryotes.

Currently, there are >8,000 genome assemblies of ~4,000 eukaryotic species in the NCBI Genome database. We downloaded 7,821 assemblies of 3,767 species and BLASTed them against the database containing human beta satDNA sequences. After filtering the BLAST outputs, we found 33,150 beta satDNA copies in 166 genome assemblies of 116 species (Fig. 1, Table 1, Supplementary Fig. S1, Supplementary Table S1, and Supplementary File 1). Since beta satDNA sequences are highly repeated and difficult to assemble into chromosome contigs, it is hard to assess the proportions of beta satDNAs in certain genomes, e.g. a human genome assembly, GCF_000002125.1_HuRef, has 7,036 beta satDNA copies, whereas, another human genome assembly, GCF_000306695.2_CHM1_1.1, has only 107 copies. To our great surprise, beta satDNAs were found in most of the major branches of eukaryotes, including 68 out of 1,394 animals, 16/415 plants, 22/1,656 fungi, 8/60 species of Apicomplexa (9/176 species of Harosa), and 1/20 species of Mycetozoa (1/43 species of Amoebozoa), yet no hit was found in Excavata (0/54) which actually is a polyphyletic group [17, 18].

Beta satDNA can be found in all the major clades of primate, as well as in tree shrew (*Tupaia chinensis*), presenting a vertical transfer scenario. In addition to primates, beta satDNAs were also found in the genome assemblies of four Bovinae species (Supplementary Table S1). HT in Bovinae had been reports in several previous studies. It seems that the genomes of Bevinae are especially prone to HT [19, 20]. Beta

satDNAs are rare in the rest of mammals, as well as in other vertebrates and invertebrates (3%), though found in most major taxa. 1.3% fungal species have been found containing beta satDNAs. Given that the fungal genomes are usually small and easy to assemble, the existence of beta satDNA in fungi might be significantly rarer than in animals. Interestingly, the proportion of genomes containing beta satDNAs is relatively higher in plants (~3.9%), and even higher in Harosa (5.1%). Moreover, the number of sequences found in Harosa is significantly higher than those in non-primate animals, fungi and plants (Table 1 and Supplementary Fig. S1). Beta satDNAs typically exist as tandem repeats in human genome, and the similar pattern were also identified in the contigs/scaffolds of certain non-primate species (Supplementary Fig. S2), indicating that they may play similar roles as they do in primate genomes. Considering that most of the current genome assemblies are far from complete, the actual distribution of beta satDNAs should be substantially wider than the current view, especially in plants, for the extra difficulty in their genome assembling [21].

Analysis of beta satDNA sequences in raw WGS data

We then analyzed 102 WGS data of 73 species from the NCBI SRA database (Fig.2 and Supplementary Table S2), so that we can 1) identify more beta satDNA sequences that have not been assembled into genome data; 2) assess the proportion of beta satDNAs in certain genomes; 3) obtain a more elaborate distribution landscape of beta satDNA by looking at some representative nodes on the tree of eukaryotes. Similar to the genome BLAST results, beta satDNAs were found in all primate clades, as well as in Dermoptera and Scandentia. Previous study suggested that there are two bursts of beta satDNA through the evolution of Hominidae: one is after the separation between great apes and lesser apes, and the other is after the separation between African great apes and orangutan [16]. However, our result showed that the proportion of beta satDNAs in orangutans is similar to those in gibbons or old world monkeys. Thus, the ‘big bang’ of beta satDNAs was taken place in the African apes, which may suggest a distinct aspect of heterochromatin regulation in the Homininae. Beta satDNAs compose >0.1% of human genome, and should receive more attentions in the coming studies.

Besides Euarchonta, bovines, toothed whales and elephants were the three mammalian clades found having beta satDNAs. Beyond mammals, beta satDNAs were identified in many parasites including blood-feeding insects. Since it is always hard to avoid human DNA contamination in isolating parasites DNA samples [22], this observation need to be treated with special caution. The existence of beta satDNAs seems more common in plants than in animals, though the abundances are not high. We found significant hits in most major taxa of plants, including Angiosperm, Gymnosperm, moss, green algae and red algae. Additionally, the three species of Harosa, *T. gondii*, *P. falciparum* and *Saccharina japonica* (brown algae), all contain beta satDNA sequences. Generally speaking, the current WGS data of protists are not sufficient and the quality is not high. We expect a better prospect of beta satDNA distribution in protists in the future.

PCR amplification of beta satDNAs in 36 species

Since there have never been reports about beta satDNAs in non-primate species or HT of beta satDNAs, our discovery is shocking. Therefore, we need fairly strong evidences to support our observations. To further validate the searching results in genome assemblies and SRAs, we performed PCR assays using genomic DNAs (gDNAs) of 36 species as templates (Fig. 3 and Supplementary Fig. S3). The positive PCR signals of beta satDNAs are typical ladders with ~70-bp spacing for their tandem repeated pattern. Basically, the PCR results were consistent with the BLAST results. The signals of mouse, rat and pig showed negative as the negative controls, while a wide range of positive signals were observed in various animal and plant genomes.

It is not surprising that the positive ratio of beta satDNA in SRA, and especially in PCR are significantly higher than that in the genome assemblies. Since beta satDNA sequences are difficult to assemble for their high repetition, many assemblies don't contain beta satDNA sequences, though the physical genomes do, e.g. we even failed to identify beta satDNA sequences from the current genome assemblies of *Pongo pygmaeus* (Bornean orangutan), who no doubt has beta satDNAs.

Of course, there are concerns of contamination in both WGS and PCR, and even the genome data have possibility of contamination and wrong assembling, hence we tried multiple approaches to rule out the risk of false conclusion introduced by possible contaminations (see Supplementary Discussion for detailed information).

Identification of beta satDNA sequences in prokaryotic genomes

To see whether beta satDNAs exist in genomes other than those of eukaryotes, we searched the 210,000 prokaryotic genome assemblies (12,398 species of 569 archaea and 11,829 bacteria) in the NCBI Genome database and found hits in 72 species, two archaea and 70 bacteria (Fig. 4). Obviously, beta satDNAs in prokaryotes are far rarer than in eukaryotes. Since the prokaryotic chromosomes are very different from the eukaryotic chromosomes and there are no centromeres or telomeres in prokaryotic chromosomes, beta satDNA may not function in prokaryotes. They could merely be results of HTs from eukaryotes. Most of the bacteria containing beta satDNAs are symbionts/pathogens of animals or plants (Supplementary Table S3) and they might acquire beta satDNAs from their hosts. In addition, there are >13,000 genome assemblies of organelles and >30,000 genome assemblies of viruses in the NCBI database, but none of them contains beta satDNA. Although viruses and symbiotic organelles are common media for HT [23, 24], they might not play roles in the HT of beta satDNAs.

Characterization of the diversity of beta satDNA sequences

To analysis the diversity and evolution of the beta satDNA sequences of different taxa, we examined the phylogenetic relationships of 1,384 beta satDNA sequences of 92 species (Fig. 5A). There is barely association between the phylogenetic tree of beta satDNAs and the tree of these species, indicating that multiple HTs have been taken place between species. However, unlike the case in coding genes or TEs, the beta satDNA sequences from different species couldn't be separated in phylogenetic tree, which may be due to the intrachromosomal and interchromosomal exchanges [25]. Moreover, the consensus or centroid sequences of beta satDNAs from different sequences are very similar too. For this reason, the HT pathways of beta satDNAs cannot be determined like those of coding sequences based on the phylogenetic relationship of sequences from different species. Then we compared the sequence libraries of different species pairwisely (Fig.5B). Clearly, the diversity of sequences in primates is higher than the rest, indicating that the pool of beta satDNA is quite small until it bursted in primates. Similarly, the cluster analysis on individual beta satDNA sequences of non-primate species showed distinct pattern from that on the total sequences (Supplementary Fig. S4 and S5).

Discussion

The unequal exchange is a strong long-range ordering force which can keep tandem arrays homogeneous [1]. The homogenization process is much faster in the bisexual species, suggesting that meiotic recombination accelerates the homogenization process [26]. Concerted evolution leads to higher homogeneity between the satDNA sequences of intraspecies than the sequences of interspecies [27, 28]. The higher-order alpha satDNAs are significantly more conservative within species than between primate species [25, 29]. However, the beta satDNA sequences lack obvious conservative property within species (Fig. 4), indicating the absence of complete concerted evolution, which might be related to a presumed reduction or suppression of meiotic recombination [27].

Employing multiple approaches, we obtained a high-resolution distribution landscape of beta satDNAs in eukaryotes. It is previously thought that beta satDNAs were originated in Catarrhini and bursted in great apes [16], whereas, our study showed that beta satDNA actually exists widely across eukaryotes. The patchy distribution of beta satDNAs across eukaryotes suggests multiple HT events during evolution. However, the existences of beta satDNA seem wider in plants, and especially in Harosa (Supplementary Table S1 and Supplementary Fig. S1). Animals, fungi and Amoebozoa belong to Opimoda, while plants and Harosa belong to Diaphoretickes. All the beta satDNA sequences identified here are within Opimoda and Diaphoretickes (Fig. 1 and 2). No full-length beta satDNAs were found in any branches of Excavata so far, with only several truncated fragments found in *T. brucei* (Euglenozoa) (data not shown). Therefore, we hypothesize that beta satDNA was an ancient sequence that originated in Diaphoretickes, after its separation with Euglenozoa. The beta satDNAs in Opimoda were diffused from Diaphoretickes via HTs. Parasites play critical roles in HT as reported previously [30, 31]. The members of Apicomplexa are all kinds of parasites of animals, including the pathogens of malaria, toxoplasmosis, cyclospriasis etc., and they might have played the major role of transferring beta satDNAs to animals. Of course, today's parasites in Apicomplexa have been co-evolving with their hosts through all the time, so that the current contents of beta satDNAs in their genomes should be the results of countless two-way HT events. The

distribution of beta satDNA in plants could be the common result of inheritance from ancestors, loss and HT. Most of the fungi that were found to have beta satDNAs are parasitic and a few saprophytic. They might have acquired beta satDNA from their hosts of the plants or animals in their living environments.

We identified beta satDNA sequences in primates, as well as in Dermoptera and Scandentia, but not in mouse, rat or other rodents. Given that the mouse genome has been thoroughly studied, we tend to believe that the mouse genome doesn't comprise beta satDNAs and beta satDNAs are absent or at least not common in rodents. Scandentia (tree shrew) was previously considered a sister group of Primatomorpha, but recently was reconsidered a sister group of Glires [32, 33]. Thus, the origin of beta satDNAs in primates could be traced back to at least the common ancestor of Primatomorpha (~80 MYA).

The studies on the repeated sequences in eukaryotic genomes are far more preliminary than the studies on regular genes, and the roles of repeated elements have long been underestimated. Several recent researches have greatly updated people's knowledge on LINE-1 retrotransposon. Ivancevic et al. reported that LINE-1 actually is widely distributed in eukaryotes and able to go through HT [34], while Percharde et al. found that LINE-1 expression is essential for the early development of mouse zygote [35]. Here for the first time, we showed the wide distribution and HT of satellite DNA. These discoveries imply that many repeated sequences, instead of being 'junk DNAs', are ancient sequences that emerged during the early evolution of eukaryotes and have been playing important roles throughout the eukaryote evolution.

HT/HGT was considered one of the major drives of evolution, while recent investigations suggest that the role of HT may be even more important than had been thought [36, 37]. Complex eukaryotic genomes usually comprise large fraction of repeated sequences, including TEs and non-TEs. Previous studies have shown that HT is important in the origin of TEs in certain genomes [20, 34, 38, 39], while here we suggest that HT may be important for the origin of satDNAs too. As the deep sequencing capacity expands tremendously, the main stream of the studies on HT will certainly transit from studies based on special cases to systematic studies [31, 34, 40, 41]. Only by this way can we evaluate the great impact of HT for evolution more accurately. We believe a new distribution landscape of beta satDNAs with higher resolution will be obtained in a few years when more genome assemblies with high quality are available.

Conclusion

We performed the largest to-date study on beta satDNAs, and for the first time, we found that beta satDNAs are widely distributed in eukaryotes, instead of existing only in primates. The beta satDNAs found in some laboratory models, such as certain species of budding yeast, fruit fly and tobacco, suggest that studies on beta satDNA can be performed in these simple organisms, besides in primate cell lines. Beta satDNAs might have originated in the ancestor of Diaphoretickes during the early evolution of eukaryotes, and the beta satDNAs in primates might have originated via a HT (or HTs) in the common ancestor of Euarchonta (~80 MYA). The previous investigations on HT were mainly focus on certain

genes or TEs. Here we showed that satellite DNAs too are materials for HT, and enriched the topics of the HT research field.

Materials And Methods

Identification of beta satDNAs from genome assemblies

All the available genomic sequences (fasta files) of eukaryotes, prokaryotes, organelles and viruses in the National Center for Biotechnology Information (NCBI) Genome Database (https://www.ncbi.nlm.nih.gov/genome/browse#!/overview/) were downloaded (Supplementary Table S4), except some assemblies that were failed to download or of low quality. The sequences were BLASTed against an index built with pre-identified beta satDNA sequences of human and other apes. The BLAST outputs were filtered using a Perl script, and sequences with match length ≥ 55 and e-Value ≤ 10^{-9} were kept. Sequences with match length = 69 and TC/GA (Sau3A restriction site) at start/end were designated as full-length beta satDNA sequences.

Identification of beta satDNAs from human chromosome Y

The full sequence of the chromosome Y (NC_000024.10) was downloaded from NCBI. The locations of beta satDNAs on chrY were determined by making preliminary comments using Geneious 11 [42], and the three regions containing beta satDNA copies, Ya, Yb and Yc were identified. Then the sequences of the three regions were mapped to a 68 bp beta satDNA reference sequence based on a 72 bp sliding window at 1 bp resolution (e.g. nt 1-72, nt 2-73…) using Geneious 11, and the dataset of the beta satDNA sequences on chrY was obtained. The scripts for extracting sequences were written in Python language.

Identification of beta satDNAs in the WGS data of different species

The raw sequencing data of the 73 species were downloaded from the NCBI SRA database (https://www.ncbi.nlm.nih.gov/sra) or BLASTed on line. Then the BLAST output files were filtered with the same criteria as the genomic BLAST outputs, and the beta satDNA sequences were extracted using scripts written in Python. The percentages of beta satDNAs in each species were calculated using the following formula:

number of hits / total number of reads in SRA databases

Generation of trees of life

The trees of life (Fig. 1, Fig. 2, Fig. 4 and Fig. S3) were originally generated at the Common Tree webpage of NCBI Taxonomy (https://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi). The .phy tree file were then modified manually and further edited at iTOL (https://itol.embl.de/) [43].

Phylogenetic analyses

Sequence alignment was performed using the MUSCLE [44], All phylogenetic analyses were conducted under maximum likelihood in RAxML [45]. The phylogenetic trees were colored at iTOL (https://itol.embl.de/).

Isolations of genomic DNAs

The larvae of *Spirometra mansoni* (plerocercoid) were collected from Chinese bull frog (*Rana rugulosa*). Briefly, frogs were anesthetized with ethanol followed by euthanasia and dissected; then the plerocercoids were isolated. Totally two frogs was used in this study. *Trypanosoma brucei* were maintained in HMI-9 medium. *Toxoplasma gondii* were isolated from cultured human fibroblast cells, as well as from mouse macrophages. Briefly, mouse infected by *T. gondii* was euthanized with CO_2 and the peritoneal fluid was extracted; then macrophages were isolated, lysed and filtered using 5 μm filter to obtain parasite cells. One mouse was used in this study. The adult *Schistosoma japonicum* were collected from rabbit. Rabbit was anesthetized using sodium pentobarbital at a dose of 100 mg/kg i.p. and euthanized with KCl 100 mg/kg i.v., and then dissected. The parasites were collected from the mesenteric veins. The total number of rabbit used is one. The genomic DNAs of the animals were isolated using TIANamp Genomic DNA Kit (TIANGEN). The genomic DNAs of plants and fungi were isolated using Plant DNA Mini Kit (OMEGA).

PCR assay

PCRs (20 μl) comprised 2 μl genomic DNA (10–80 ng), 2× Premix Taq (TaKaRa), 0.5 μl forward / reverse primers (see below for sequences). The PCR cycling program was 30 cycles of 98 °C for 10 s, 56 °C for 30 s, 72 °C for 1 min and a final extension step of 72 °C for 10 min. The PCR products were checked by electrophoresis with 1.5% TAE agarose gel containing gel-red.

Primer F: GATCACCCAGGTGATGTAACTCTTGTC

Primer R: GATCAGTGCAGAGATATGTCACAATGCC

Next generation sequencing of amplicons

Barcoded primers were used in PCR for amplifying beta satDNAs from different gDNA samples. The amplicons were purified and sequenced at Guangzhou IGE Biotechnology. The raw sequence reads were then BLASTed against the beta satDNA index and filtered under the same parameter as the genome and SRA BLAST filtering.

Cluster analysis of beta satDNA sequences

Full-length beta satDNA sequences were compared pairwisely. The matrix of identities was used for the clustering analysis and heatmap generation using the heatmap.2 function of 'gplots' library in R language.

Cluster analysis of sequences identified from different genome assemblies

For the genome assemblies containing beta satDNA sequences, those that contain ≥10 beta satDNA copies were chosen. The beta satDNA libraries were compared pairwisely and the common ratios were calculated using the algorithm as below:

R-common = (libA ∩ libB)/(libA ∪ libB)

And a matrix of R-common was generated. Considering the sizes if libraries varies greatly, a matched random control matrix was created by randomizing the sequences and ran the script for 20 times to get an average. Then the difference of the two matrixes was used for the clustering analysis and heatmap generation using the heatmap.2 function in R language. GCF_000306695.2_CHM1_1.1 and GCA_002754635.1_CMB-1_v2 were excluded in Fig. 5B for very different copy numbers in relative genera.

Abbreviations

satDNA, satellite DNA; HT, horizontal transfer; HGT, horizontal gene transfer; WGS, whole genome sequencing; SRA, short reads archive; TE, transposable element.

Declarations

Ethics approval and consent to participate

The animal experiments were supervised by the Institutional Animal Care and Use Committee of the Sun Yat-sen University (IACUC-SYSU) and performed in accordance with the regulation and guidelines of this committee (Approval No. SYSU-IACUC-2019-B048).

Consent for publication

Not applicable

Availability of data and materials

Please see supplementary file 1 for the beta satDNA sequences.

Competing interest: The authors declare that they have no competing interests.

Funding: This work was supported by National Natural Science Foundation of China (81872295 to Y. G.); Guangdong Natural Science Foundation (2018A030313819 to Y. G.); Guangdong Science and Technology Department (2017B030314026).

Authors' contributions

J.Y. and Y.G. designed the project. Y.W., M.L. and J.L. collected the parasites. B.Y., D.X., J.Y. and Y.G. collected the plant and fungi samples. J.Y., G.M., Y.Z., Y.Z., and Z.H.G. isolated the genomic samples and did PCR. J.Y., Z.H.G. and Y.G. arranged data, wrote scripts and performed bioinformatics analyses. J.Y. and Y.G. wrote the manuscript.

Acknowledgement

We thank Dr. Henry L. Levin of NICHD, NIH for his valuable advices.

References

1. Charlesworth B, Sniegowski P, Stephan W: **The evolutionary dynamics of repetitive DNA in eukaryotes.** *Nature* 1994, **371:**215-220.
2. Yunis JJ, Yasmineh WG: **Heterochromatin, satellite DNA, and cell function. Structural DNA of eucaryotes may support and protect genes and aid in speciation.** *Science* 1971, **174:**1200-1209.
3. Greig GM, Willard HF: **Beta satellite DNA: characterization and localization of two subfamilies from the distal and proximal short arms of the human acrocentric chromosomes.** *Genomics* 1992, **12:**573-580.

4. Khost DE, Eickbush DG, Larracuente AM: **Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in Drosophila melanogaster.** *Genome Res* 2017, **27:**709-721.
5. Tomilin NV: **Regulation of mammalian gene expression by retroelements and non-coding tandem repeats.** *Bioessays* 2008, **30:**338-348.
6. Mirkin SM: **DNA structures, repeat expansions and human hereditary disorders.** *Curr Opin Struct Biol* 2006, **16:**351-358.
7. Mirkin SM: **Expandable DNA repeats and human disease.** *Nature* 2007, **447:**932-940.
8. J. Rich, V. V. Ogryzko, Pirozhkova IV: **Satellite DNA and related diseases.** *Biopolymers and Cell* 2014, **30:**249–259.
9. Scott HS, Kudoh J, Wattenhofer M, Shibuya K, Berry A, Chrast R, Guipponi M, Wang J, Kawasaki K, Asakawa S, et al: **Insertion of beta-satellite repeats identifies a transmembrane protease causing both congenital and childhood onset autosomal recessive deafness.** *Nat Genet* 2001, **27:**59-63.
10. Zhu Q, Pao GM, Huynh AM, Suh H, Tonnu N, Nederlof PM, Gage FH, Verma IM: **BRCA1 tumour suppression occurs via heterochromatin-mediated silencing.** *Nature* 2011, **477:**179-184.
11. Meneveri R, Agresti A, Della Valle G, Talarico D, Siccardi AG, Ginelli E: **Identification of a human clustered G + C-rich DNA family of repeats (Sau3A family).** *J Mol Biol* 1985, **186:**483-489.
12. Meneveri R, Agresti A, Marozzi A, Saccone S, Rocchi M, Archidiacono N, Corneo G, Della Valle G, Ginelli E: **Molecular organization and chromosomal location of human GC-rich heterochromatic blocks.** *Gene* 1993, **123:**227-234.
13. Agresti A, Rainaldi G, Lobbiani A, Magnani I, Di Lernia R, Meneveri R, Siccardi AG, Ginelli E: **Chromosomal location by in situ hybridization of the human Sau3A family of DNA repeats.** *Hum Genet* 1987, **75:**326-332.
14. Eichler EE, Hoffman SM, Adamson AA, Gordon LA, McCready P, Lamerdin JE, Mohrenweiser HW: **Complex beta-satellite repeat structures and the expansion of the zinc finger gene cluster in 19p12.** *Genome Res* 1998, **8:**791-808.
15. Meneveri R, Agresti A, Rocchi M, Marozzi A, Ginelli E: **Analysis of GC-rich repetitive nucleotide sequences in great apes.** *J Mol Evol* 1995, **40:**405-412.
16. Cardone MF, Ballarati L, Ventura M, Rocchi M, Marozzi A, Ginelli E, Meneveri R: **Evolution of beta satellite DNA sequences: evidence for duplication-mediated repeat amplification and spreading.** *Mol Biol Evol* 2004, **21:**1792-1799.
17. Cavalier-Smith T, Chao EE, Lewis R: **Multiple origins of Heliozoa from flagellate ancestors: New cryptist subphylum Corbihelia, superclass Corbistoma, and monophyly of Haptista, Cryptista, Hacrobia and Chromista.** *Mol Phylogenet Evol* 2015, **93:**331-362.
18. Brown MW, Heiss AA, Kamikawa R, Inagaki Y, Yabuki A, Tice AK, Shiratori T, Ishida KI, Hashimoto T, Simpson AGB, Roger AJ: **Phylogenomics Places Orphan Protistan Lineages in a Novel Eukaryotic Super-Group.** *Genome Biol Evol* 2018, **10:**427-433.

19. Kordis D, Gubensek F: **Horizontal SINE transfer between vertebrate classes.** *Nat Genet* 1995, **10:**131-132.
20. Walsh AM, Kortschak RD, Gardner MG, Bertozzi T, Adelson DL: **Widespread horizontal transfer of retrotransposons.** *Proc Natl Acad Sci U S A* 2013, **110:**1012-1016.
21. Claros MG, Bautista R, Guerrero-Fernández D, Benzerki H, Seoane P, Fernández-Pozo N: **Why Assembling Plant Genome Sequences Is So Challenging.** *Biology* 2012, **1:**439-459.
22. Oyola SO, Gu Y, Manske M, Otto TD, O'Brien J, Alcock D, Macinnis B, Berriman M, Newbold CI, Kwiatkowski DP, et al: **Efficient depletion of host DNA contamination in malaria clinical sequencing.** *J Clin Microbiol* 2013, **51:**745-751.
23. Bergthorsson U, Adams KL, Thomason B, Palmer JD: **Widespread horizontal transfer of mitochondrial genes in flowering plants.** *Nature* 2003, **424:**197-201.
24. Liu H, Fu Y, Jiang D, Li G, Xie J, Cheng J, Peng Y, Ghabrial SA, Yi X: **Widespread horizontal gene transfer from double-stranded RNA viruses to eukaryotic nuclear genomes.** *J Virol* 2010, **84:**11876-11887.
25. Rudd MK, Wray GA, Willard HF: **The evolutionary dynamics of alpha-satellite.** *Genome Res* 2006, **16:**88-96.
26. Mantovani B, Tinti F, Bachmann L, Scali V: **The Bag320 satellite DNA family in Bacillus stick insects (Phasmatodea): different rates of molecular evolution of highly repetitive DNA in bisexual and parthenogenic taxa.** *Mol Biol Evol* 1997, **14:**1197-1205.
27. Kuhn GC, Franco FF, Manfrin MH, Moreira-Filho O, Sene FM: **Low rates of homogenization of the DBC-150 satellite DNA family restricted to a single pair of microchromosomes in species from the Drosophila buzzatii cluster.** *Chromosome Res* 2007, **15:**457-469.
28. Dover G: **Molecular drive: a cohesive mode of species evolution.** *Nature* 1982, **299:**111-117.
29. Cacheux L, Ponger L, Gerbault-Seureau M, Richard FA, Escude C: **Diversity and distribution of alpha satellite DNA in the genome of an Old World monkey: Cercopithecus solatus.** *BMC Genomics* 2016, **17:**916.
30. Deitsch K, Driskill C, Wellems T: **Transformation of malaria parasites by the spontaneous uptake and expression of DNA from human erythrocytes.** *Nucleic Acids Res* 2001, **29:**850-853.
31. Gilbert C, Schaack S, Pace JK, 2nd, Brindley PJ, Feschotte C: **A role for host-parasite interactions in the horizontal transfer of transposons across phyla.** *Nature* 2010, **464:**1347-1350.
32. Meredith RW, Janecka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik E, Simao TL, Stadler T, et al: **Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification.** *Science* 2011, **334:**521-524.
33. Zhou X, Sun F, Xu S, Yang G, Li M: **The position of tree shrews in the mammalian tree: Comparing multi-gene analyses with phylogenomic results leaves monophyly of Euarchonta doubtful.** *Integr Zool* 2015, **10:**186-198.

34. Ivancevic AM, Kortschak RD, Bertozzi T, Adelson DL: **Horizontal transfer of BovB and L1 retrotransposons in eukaryotes.** *Genome Biol* 2018, **19:**85.
35. Percharde M, Lin CJ, Yin Y, Guan J, Peixoto GA, Bulut-Karslioglu A, Biechele S, Huang B, Shen X, Ramalho-Santos M: **A LINE1-Nucleolin Partnership Regulates Early Development and ESC Identity.** *Cell* 2018, **174:**391-405.
36. Keeling PJ, Palmer JD: **Horizontal gene transfer in eukaryotic evolution.** *Nat Rev Genet* 2008, **9:**605-618.
37. Daubin V, Szollosi GJ: **Horizontal Gene Transfer and the History of Life.** *Cold Spring Harb Perspect Biol* 2016, **8:**a018036.
38. Gilbert C, Feschotte C: **Horizontal acquisition of transposable elements and viral sequences: patterns and consequences.** *Curr Opin Genet Dev* 2018, **49:**15-24.
39. Gilbert C, Hernandez SS, Flores-Benabib J, Smith EN, Feschotte C: **Rampant horizontal transfer of SPIN transposons in squamate reptiles.** *Mol Biol Evol* 2012, **29:**503-515.
40. Ivancevic AM, Kortschak RD, Bertozzi T, Adelson DL: **LINEs between Species: Evolutionary Dynamics of LINE-1 Retrotransposons across the Eukaryotic Tree of Life.** *Genome Biol Evol* 2016, **8:**3301-3322.
41. Peccoud J, Loiseau V, Cordaux R, Gilbert C: **Massive horizontal transfer of transposable elements in insects.** *Proc Natl Acad Sci U S A* 2017, **114:**4721-4726.
42. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, et al: **Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data.** *Bioinformatics* 2012, **28:**1647-1649.
43. Letunic I, Bork P: **Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees.** *Nucleic Acids Res* 2016, **44:**W242-245.
44. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32:**1792-1797.
45. Stamatakis A: **RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.** *Bioinformatics* 2014, **30:**1312-1313.

Tables

Table 1. A summary of beta satDNA sequences found in the genome assemblies of eukaryotes.

Taxa	Species	copy number	copy number of full-length units
primate	28	29067	16140
animal excluding primate	40	673	445
fungus	22	317	252
Amoebozoa	1	19	17
plant	16	435	321
Harosa	9	2639	2103
total	116	33150	19278
unique seqences		18390	10966

Figures

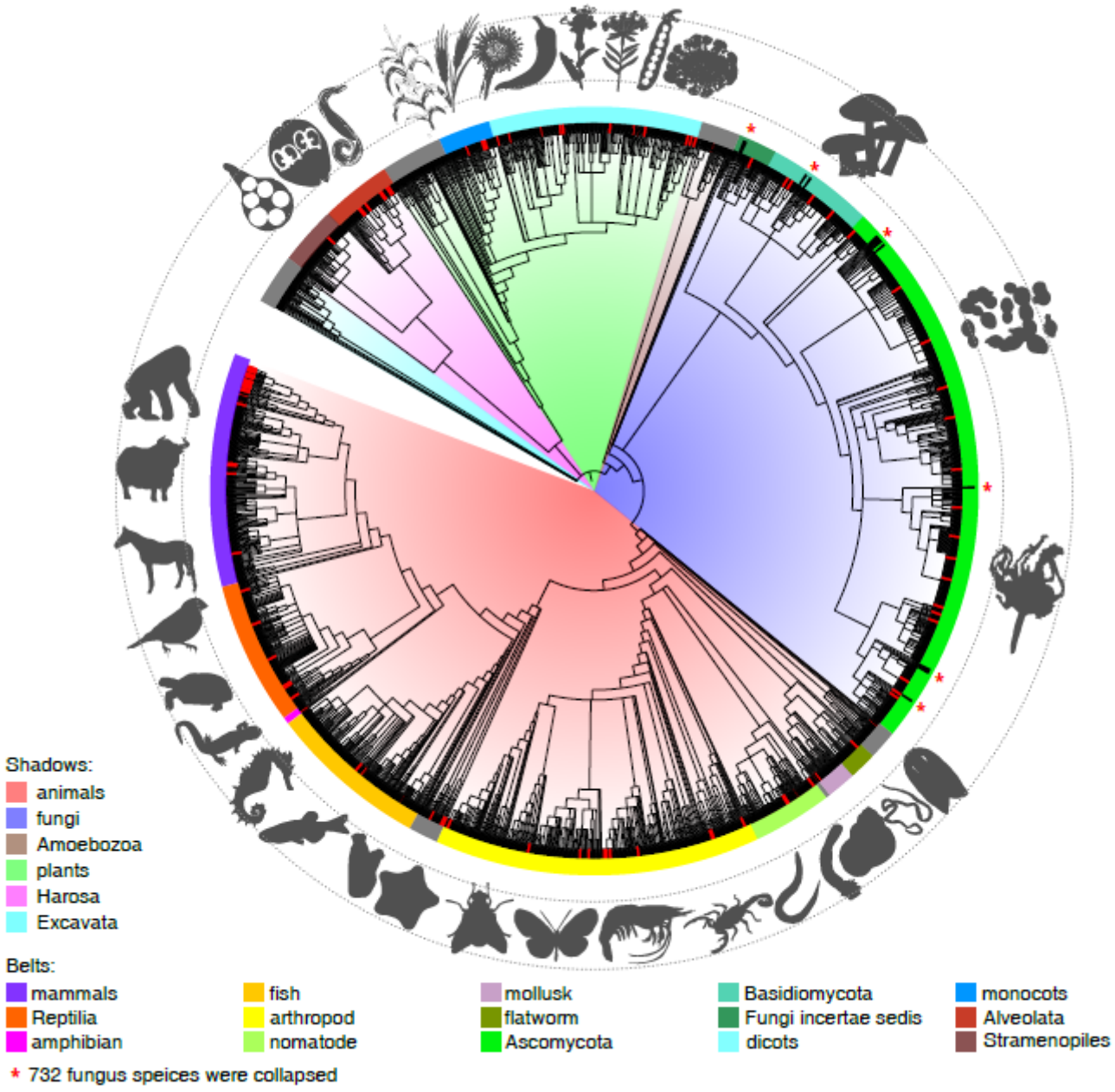

Figure 1

The distribution landscape of beta satDNAs in eukaryotes. 7,821 genome assemblies of 3,767 eukaryotic species were downloaded from the NCBI Genome database and BLASTed against the index of beta satDNA. 116 species (red lines) were found containing beta satDNA sequences.

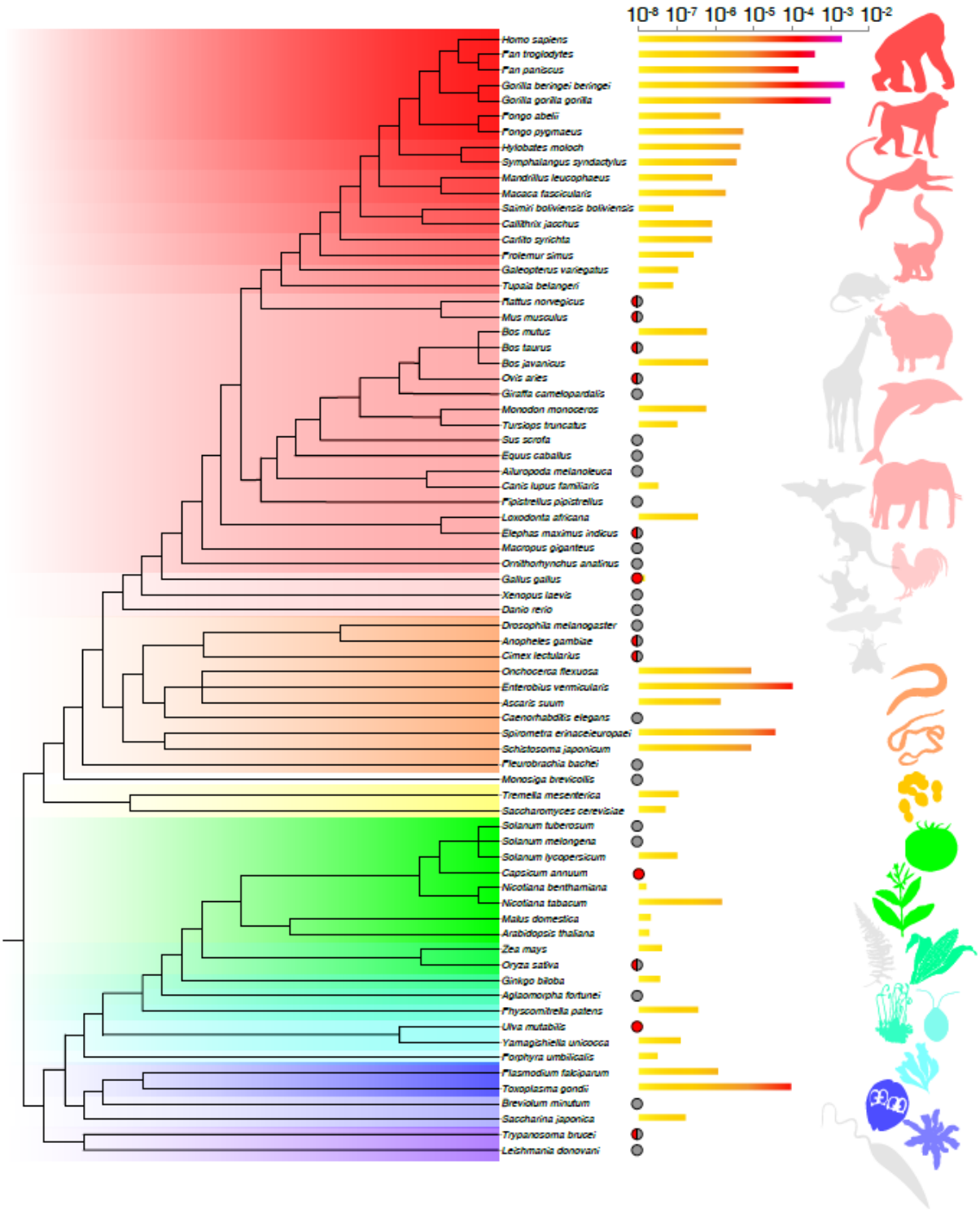

Figure 2

Beta satDNAs in WGS data (SRA). 102 SRA datasets of 73 species were BLASTed against the beta satDNA index. The proportions of beta satDNAs in different species were shown by the histogram at the right part. The evolutionary relationships of the species were shown at the left part. Color range: red, vertebrates (from light to dark: vertebrates, mammals, Euarchonta, primates, Simiiformes, Catarrhini, apes and Hominidae); orange, invertebrates; cyan-green, plants (from cyan to green: plants, green plants,

Embryophytes, Equisetophyta, Equisetophytina, Spermatophytes, Angiosperms, dicots); blue, Harosa (dark, Alveolata); purple, Euglenozoa. Circles: gray, not found or ratio <10-8; red, found but the bars are too short to view; half red half gray, found in certain SRAs but not found in others.

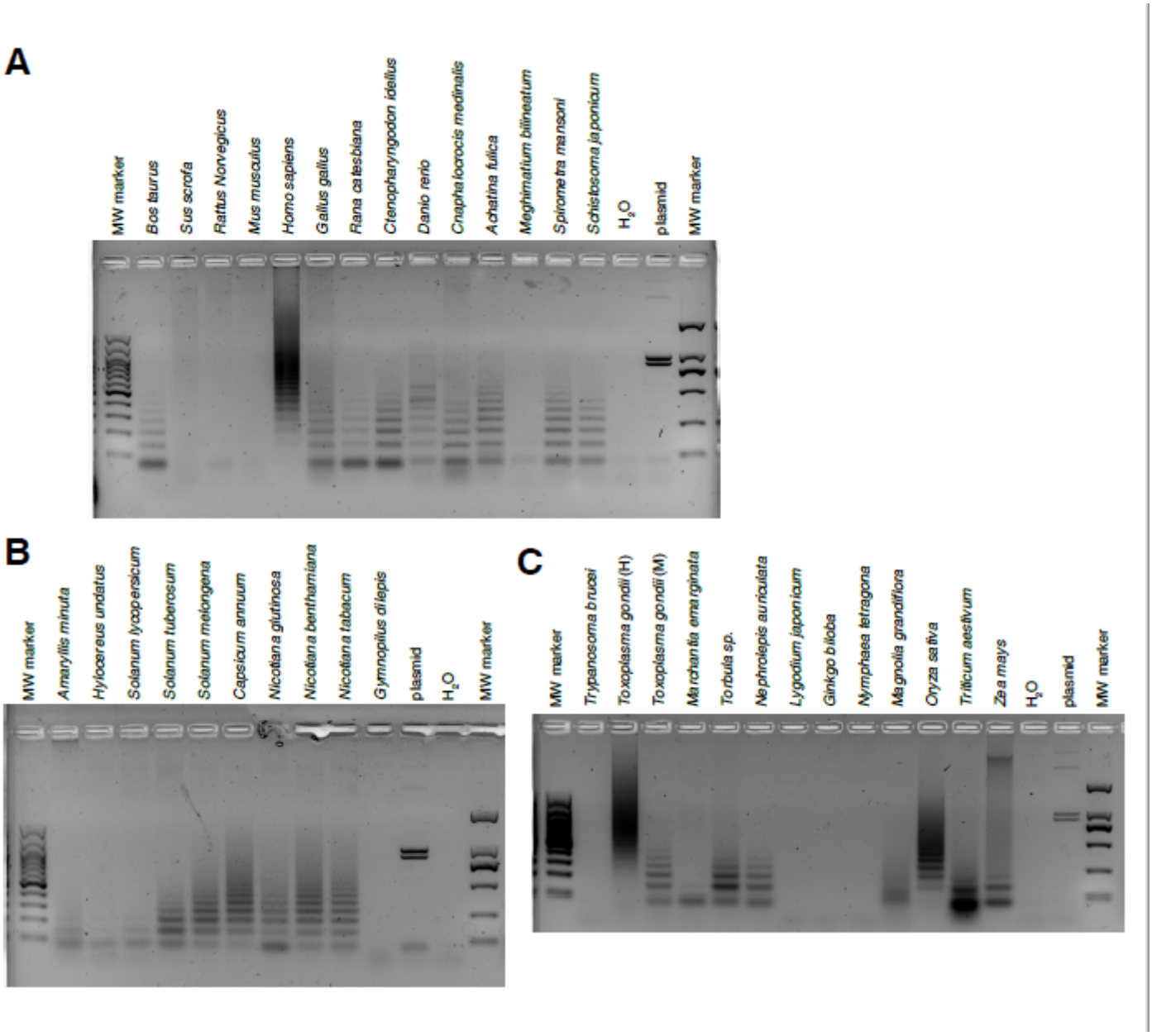

Figure 3

PCR amplifications of beta satDNAs. Genomic DNA samples were isolated from certain organisms and used as templates. Primers were designed according to the consensus sequence of beta satDNA. The products were then analyzed using electrophoresis with 1.5% agarose gel containing gel-red.

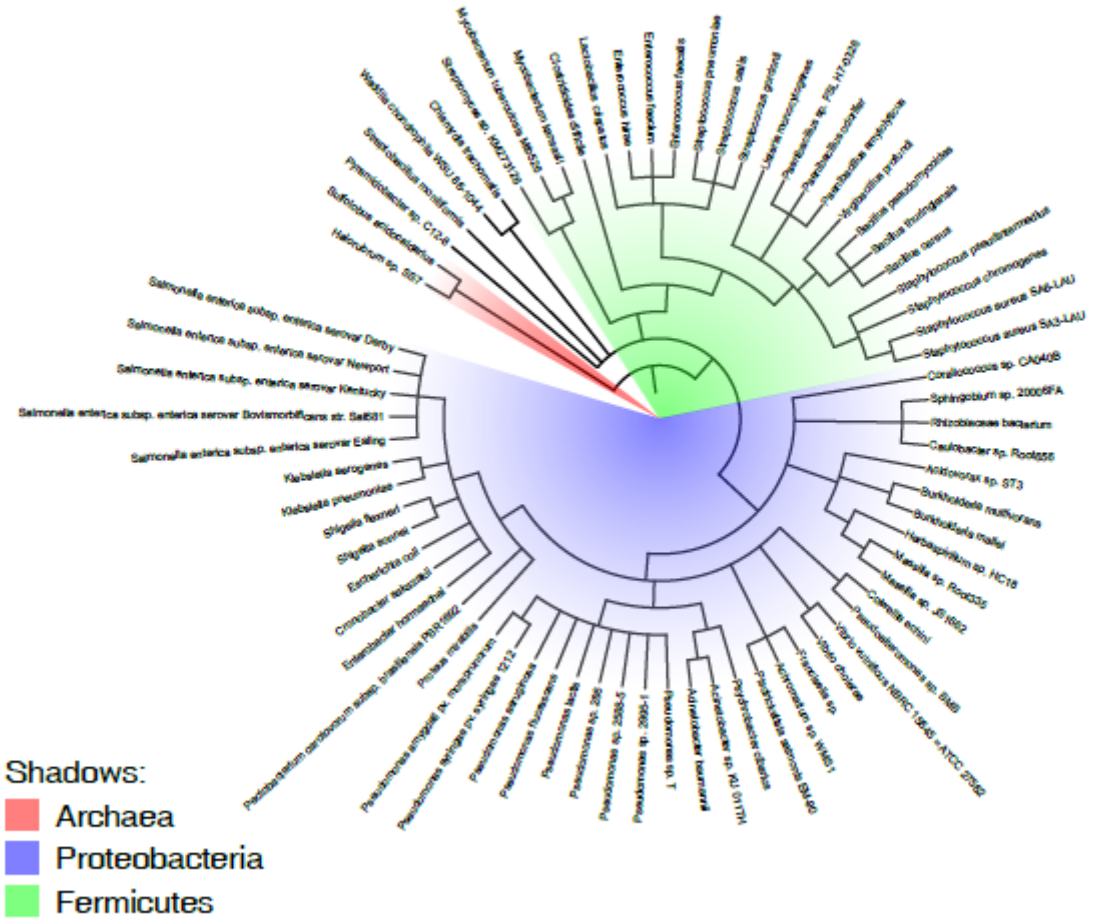

Figure 4

Prokaryotic species containing beta satDNA sequences and their relationship. >210,000 prokaryotic genome assemblies were BLASTed against the beta satDNA index and 72 species were found containing beta satDNA sequences.

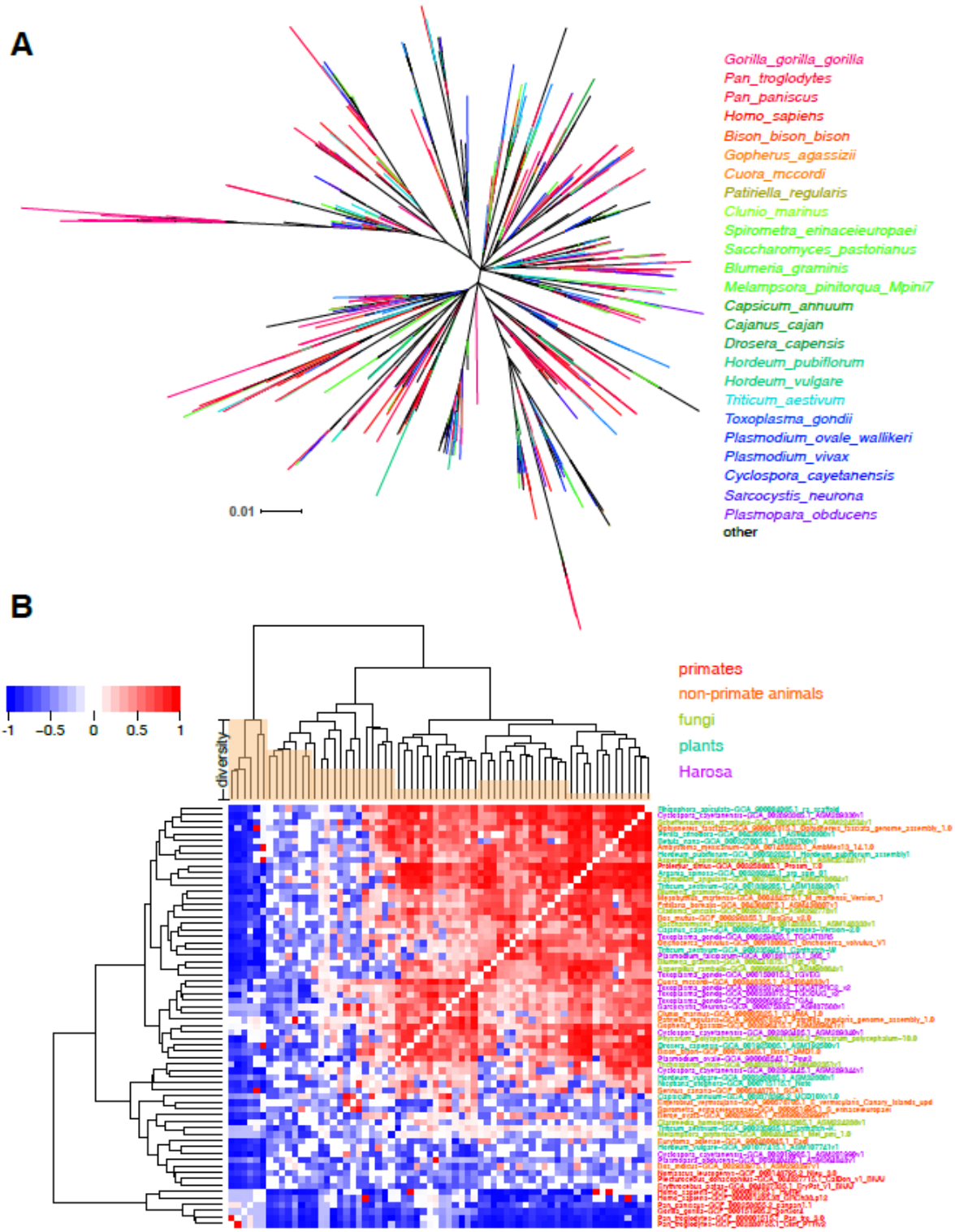

Figure 5

Prokaryotic species containing beta satDNA sequences and their relationship. >210,000 prokaryotic genome assemblies were BLASTed against the beta satDNA index and 72 species were found containing beta satDNA sequences.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SuppTable4.xlsx
- SupplementaryMaterials.pdf
- betaSatDNAseq.fasta
- SuppTable13.xlsx