

RESEARCH

Open Access



# SINE-derived satellites in scaled reptiles

Nikita S. Vassetzky<sup>1,2\*</sup>, Sergei A. Kosushkin<sup>2</sup> and Alexey P. Ryskov<sup>1</sup>

## Abstract

**Background** The genomes of many eukaryotes contain DNA repeats in the form of both tandem and interspersed elements with distinct structure, evolutionary histories, and mechanisms of emergence and amplification. Although there is considerable knowledge regarding their diversity, there is little evidence directly linking these two types.

**Results** Different tandem repeats derived from portions of short interspersed elements (SINEs) belonging to different families were identified in 56 genomes of squamate reptiles. All loci of SINE-derived satellites (sSats) were thoroughly analyzed. Snake sSats exhibited high similarity in both structure and copy number, while other taxa may have highly diverse (geckos), rare (*Darevskia* lizards), or missing sSats (agamid lizards). Similar to most satellites associated with heterochromatin, sSats are likely linked to subtelomeric chromosomal regions.

**Conclusions** Discovered tandem repeats derived from SINEs exhibit satellite-like properties, although they have not amplified to the same degree as typical satellites. The autonomous emergence of distinct sSats from diverse SINE families in numerous squamate species suggests a nonrandom process of satellite genesis originating from repetitive SINEs.

**Keywords** Satellite DNA, Tandem repeats, SINEs, Retrotransposons, Squamata, Reptilia, Evolution

## Background

Eukaryotic genomes contain two primary types of repetitive elements: tandem repeats and transposable (interspersed) elements. These elements are highly diverse, comprising many classes within a genome and differing in structure, genomic organization, and amplification mechanisms.

Primarily, tandem repeats are organized in long arrays of relatively short non-coding sequences called satellites. Satellites are typically categorized into two or three classes: micro-, mini- and regular satellites. However, sometimes minisatellites are placed in both the

micro- and satellite categories. We will follow the division into microsatellites (simple repeats) and proper satellites [1].

Satellite DNAs exhibit variations in nucleotide sequence, sequence complexity, repeat unit length, and abundance. However, they do share two essential features: organization into lengthy arrays of tandem head-to-tail repeats and association with heterochromatic (telomeric or centromeric) regions [2]. The content of satellite DNA varies from 0.5% to more than 50% in animal genomes [1].

Different genomes may contain satellite families specific to certain species, while others may be shared across numerous taxa. However, most satellite families within a species have unrelated sequences. For example, the human genome contains nine satellite families with the predominant  $\alpha$ -satellite comprising over half of total satellite DNA [1].

Structural roles of satellite DNA in chromosome organization, pairing, and segregation have been proposed. Telomeric and centromeric regions stabilize the

\*Correspondence:

Nikita S. Vassetzky  
nvas@eimb.ru

<sup>1</sup> Institute of Gene Biology of the Russian Academy of Sciences,  
Moscow 119334, Russia

<sup>2</sup> Engelhardt Institute of Molecular Biology, Russian Academy of Sciences,  
Moscow 119991, Russia



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

chromatin at these sites for effective interactions with DNA-binding proteins, which is crucial for kinetochore formation and chromosomal segregation during mitosis and meiosis. Subtelomeric satellites stabilize chromosomal ends in a sequence-independent manner [1].

A different type of DNA repeats are transposable elements that comprise two classes: retrotransposons and DNA transposons. Retrotransposons utilize RNA-mediated mechanisms and a copy-and-paste process in their amplification, while DNA transposons rely on DNA-mediated mechanisms and a cut-and-paste process.

Short interspersed elements (SINEs) and long interspersed elements (LINEs) are the most abundant repetitive elements in higher eukaryotes, with SINEs outnumbering LINEs in most vertebrates and plants [3]. SINEs do not encode any proteins and their amplification is dependent on the enzymes of the cell and the partner LINEs. SINEs originate from a limited number of “master” copies, which can vary over time and give rise to numerous subfamilies [4]. SINEs are predominantly found in euchromatin [5].

Certain retrotransposons have distinct evolutionary relationships, such as the 3′-terminal region of a typical SINE originating from its partner LINE. Some SINE families, mostly in mammals, include (TC)<sub>n</sub> stretches, which resemble microsatellites in both structure and behavior; these structures are hypervariable sites within SINEs [4]. The structure of a SINE typically comprises four distinct regions: a head derived from one of three types of cellular RNAs transcribed by RNA polymerase III (most commonly, tRNA); a body whose origin and function remain largely unknown; a LINE-derived region (LDR) the region necessary for recognition of SINE RNA by the LINE machinery; and a tail composed of a variable-length sequence of simple repeats. This pattern is applicable to three SINE families, Sauria/Squam1 (here referred to as Squam1 to avoid confusion with the Sauria taxonomic name), Squam2, and Squam3, which are relevant to this study [4].

Despite their differing structures and amplification mechanisms, transposable elements and satellites can share similar sequences, suggesting evolutionary relationships. Segments of transposons can be amplified as satellites as reported for LTR retrotransposons, LINEs, SINEs, and DNA transposons. We analyzed a limited number of published SINE-satellite examples and found them inconclusive. Please refer to specific reviews for information about other TE origins of satellites [6, 7].

- (1) The PolIII/TAN (formerly known as Hirt) tandem repetitive family in the newt *Cynops pyrrhogaster* [8] contains a tRNA-like sequence and is capable of in vitro transcription. Although this ~360-nt tan-

dem repeat does include a tRNA-related sequence, no similar SINEs have been identified. This species is underrepresented in sequence databases, so we cannot confirm the abundance of these tandem repeats. Thus, it can hardly represent the SINE-satellite relationship.

- (2) The OAX repeat in *Xenopus* has a central segment derived from tRNA and is claimed to be derived from a SINE [9]. Although this region likely originates from tRNA, its relation to a *Xenopus* SINE is yet to be confirmed. Even though these tandem repeats were designated as ‘satellite I’ [10], their repeat unit is too long to be considered a proper satellite.
- (3) The Rana/polIII is a highly repetitive element, accounting for up to 10% of the genomes of certain Palearctic green frogs (*Rana esculenta* group) [11]. The ~250 nt tandem of this element shows some similarity with tRNA (64%/64 nt) but no similarity with any known SINEs outside of this particular region.
- (4) Finally, the European salamanders *Hydromantes* have a Hy/polIII repetitive family with a ~200-nt tandem repeat [12]. However, even the origin of this sequence from tRNA is questionable (there is no box B), let alone its assignment to SINEs.

When studying Squam1 and Squam3 SINEs in reptiles [13], irregular copies were found with a few or multiple tandem duplications within a SINE sequence resembling satellites. Here, we demonstrated that the association between SINEs and tandem repeats is not occasional (at least in scaled reptiles) and analyzed different aspects of this association for three SINE families (Squam1, Squam2, and Squam3) in a variety of full-genome assemblies of scaled reptiles (Squamata). These SINEs are tRNA-related but have independent origin [13–15].

## Results

Initially, we identified repetitive elements with tandem duplications of SINE-derived sequences in the Indian cobra *Naja naja* and other species in the study of Squam3 SINEs [13]. We then conducted a systematic search for similar elements across 56 completely sequenced genomes using diverse bioinformatic approaches, including some that were custom-made for this purpose. Such SINE-derived satellites (sSat) were discovered for three SINE families, Sauria (Squam1) [14, 15], Squam2 [16], and Squam3 [13], and are referred to as sSat1, sSat2, and sSat3, respectively.

The sequence lengths of the three squamate SINE families range from approximately 190 to 350 nt. Notably, the sSat repeat units were around 70 to 290 nt but

always shorter than the corresponding SINE, and they could cover the 5′-, middle, and/or 3′-parts of SINE (Fig. 1). The number of tandem repeat units at a specific location varied greatly, ranging from a few to hundreds. SSats with increased numbers of repeat units exhibited a uniform configuration, consistent with satellite formations. Conversely, sSats with fewer repeat units displayed greater variability in both the corresponding SINE region and sequence. The former (with at least four repeat units) were considered as nuclei for the proper satellites and given closer attention.

In our analysis of tandem sSat, the terminal 5′ (leading) and 3′ (trailing) monomers were distinguished from the internal (middle) ones. We largely considered sSat loci with at least four middle monomers (unless stated otherwise).

### Squam3-derived satellites

(sSat3) were found to be plentiful in squamates. While certain taxa lack or have marginal levels of sSat loci (disregarding Squam3 SINE numbers), five squamate taxa are known to have sSats: snakes (Serpentes), geckonids (Gekkota), teiids (Teiidae), iguanas (Iguania), and spectacled lizards (Gymnophthalmidae).

### Squam3-derived satellites in snakes (Serpentes)

The great majority of sSats tested across 19 snake genomes exhibit high structural and sequence similarity. The tandem repeat unit covers a 3′-portion of Squam3 head, which includes the promoter box B, and a section of the SINE body (Fig. 1). The unit spans 67 nt and corresponds to positions 42–108 in Squam3C, the snake-specific Squam3 subfamily. However, two alternative tandem variants are observed in boas (Supplementary Fig. 1). The python, another primitive snake, only had 15 sSat3 loci with few tandem repeat units (mostly 2–4 monomers, at most 8) corresponding to different parts of Squam3. This is the only such example of sSat3 found in snakes; however, similar low-copy-number SINE-derived satellites were not uncommon in other squamates. We provide more detailed information on this matter in Supplementary Fig. 2.

Additionally, apart from this major tandem repeat, we encountered other variants, which contained a lesser

number of repeat units and fewer loci. During our analysis, we focused on the major variants.

There was a high average similarity in the sequences found within and between sSat3 loci. As an illustration, in a sample of 45 Indian cobra loci the average sequence similarity of tandem repeat units was 89%, while the average mean similarity of Squam3 SINE sequences was 64% [13]. Meanwhile, longer loci exhibited greater homogeneity in their repeat units. For instance, the mean similarity was 82, 87, and 91% in the Indian cobra loci including 4, 5–9, and more than 10 repeat units, respectively; and the longest locus featuring 320 monomers had a similarity rate as high as 94%.

In some instances of sSat3 loci, the tandem repeat units can be separated into two or even three locus-specific subvariants characterized by a particular set of substitutions and indels. Additionally, these subvariants are frequently alternated. A manual examination of sSat3 sequences in the Indian cobra and glossy snake *Arizona elegans* showed this alternation occurring in up to 25% of lengthy loci (with a minimum of 20 repeat units) (Supplementary Fig. 3).

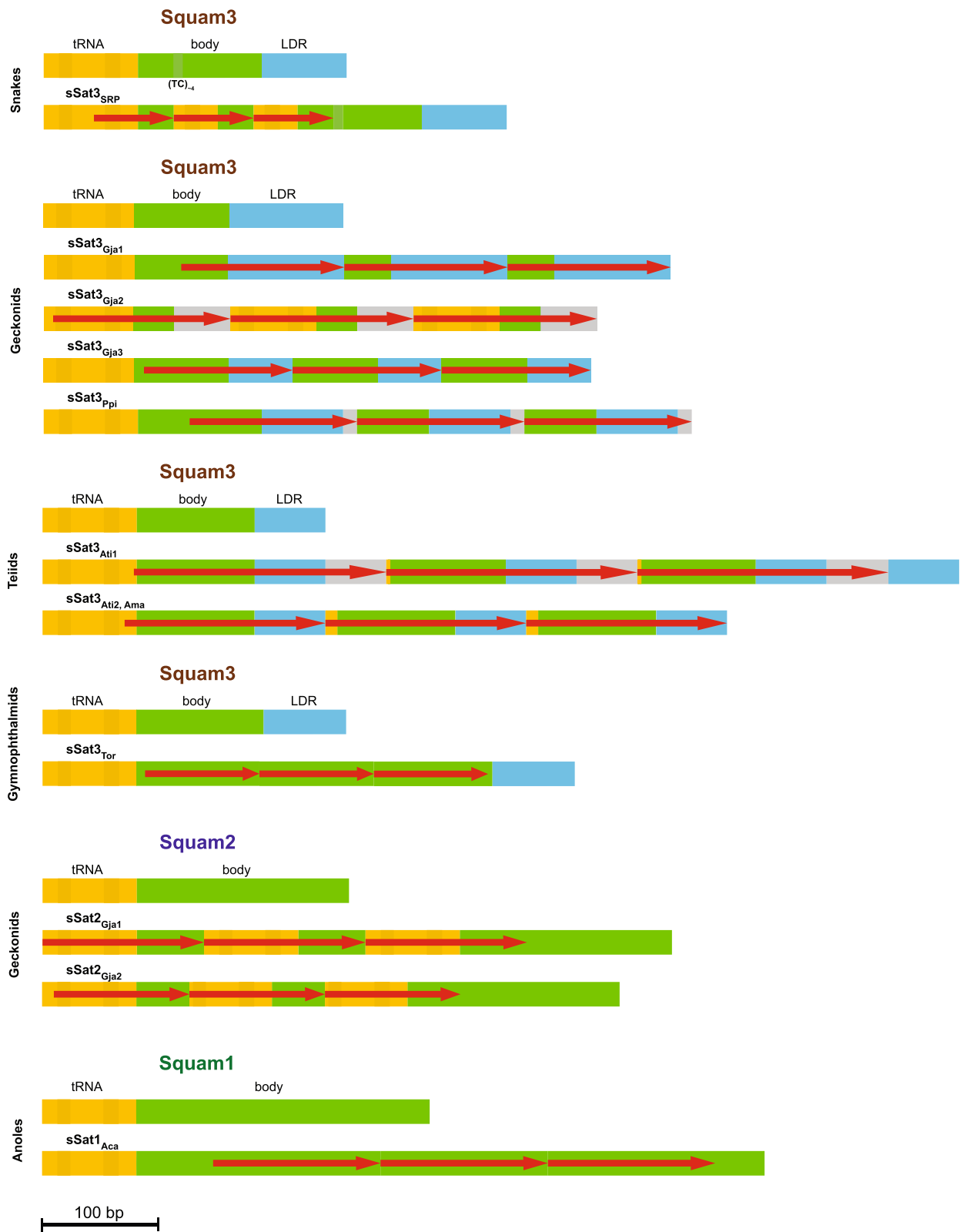
The presence of sSat3 loci, which can number in the thousands in most snake species, may be less prevalent (in the hundreds and tens) in certain species or completely absent in the blind snake *Aniliios bituberculatus* (Fig. 2). Typically, the number of sSat3 loci correlated with the number of Squam3 SINE; however there are a significant number of exceptions to this correlation.

Regarding the number of tandem repeat units in sSat3 loci, most snakes exhibit a relatively high proportions of lengthy sSat stretches. For instance, the loci of the Indian cobra and glossy snake contain up to 320 and nearly 950 repeat units, respectively. It is important to note that these data, particularly for the longest loci, are significantly impacted by the quality of genome assembly. Consequently, the number of sSat loci may significantly differ within the same genus (Fig. 2).

Snakes exhibit impressive consistency in the taxonomic distribution of sSat3, with no such satellite sequences found exclusively in the blind snake *Aniliios bituberculatus*. The Burmese python *Python bivittatus* had as few as 15 sSat3 loci. The *Boa constrictor* had a relatively low number of sSat3 loci, while another boa species *Charina*

(See figure on next page.)

**Fig. 1** Schematic diagram of major SINE-derived satellites in squamate genomes. SINE (Squam3, Squam2, and Squam1) structure includes the tRNA-related head with the shaded promoter boxes (orange), body (green), and LINE-derived region (blue); the respective sSat regions are colored the same. Regions marked in gray have neither relation to known sequences nor to each other. The red arrows mark the tandem repeat unit (their numbers varied widely). The original SINE is specified above each panel for a given taxon (on the left). The subscript indices mark the corresponding taxon or species: SRP, Serpentes; Gja, *Gekko japonicus*; Ppi, *Paroedura picta*; Ati, *Aspidoscelis tigris*; Ama, *Aspidoscelis marmoratus*; Tor, *Tretioscincus oriximinensis*; and Aca, *Anolis carolinensis*



**Fig. 1** (See legend on previous page.)

		Genome size, Gbp	sSat1					sSat2					sSat3					
			Squam1 SINE	loci				Squam2 SINE	loci				Squam3 SINE	loci				
				total	4-5	6-8	≥9		total	4-5	6-8	≥9		total	4-5	6-8	≥9	
Gekkota	Gekkonidae	<i>Gekko japonicus</i>	2.9	70338	40	38	2	0	62287	2580	1620	761	199	229108	6039	5040	884	115
		<i>Paroedura picta</i>	1.8	0	0			23024	262	227	34	1	140016	2755	2621	134	0	
	Sphaerodactylidae	<i>Sphaerodactylus townsendi</i>	1.9	0	0			22548	38	21	14	3	88633	404	365	31	8	
	Eublepharidae	<i>Eublepharis macularius</i>	2.2	0	0			38389	40	29	7	4	108054	61	41	9	11	
Lacertoidea	Darevskia	<i>Darevskia valentini</i>	2.0	77377	12	12	0	0	25283	38	37	1	0	41927	0			
		<i>Darevskia raddei</i>	2.0	64152	11	11	0	0	23666	35	35	0	0	35761	0			
		<i>Darevskia unisexualis</i>	2.0	69440	9	9	0	0	28712	35	33	2	0	40399	0			
		<i>Lacerta agilis</i>	1.8	48807	11	11	0	0	24954	38	35	2	1	42980	2	2	0	0
	Lacertidae	<i>Lacerta viridis</i>	1.9	50632	16	15	1	0	25667	34	32	2	0	39683	6	1	0	5
		<i>Lacerta bilineata</i>	1.8	50838	9	8	0	1	25210	30	29	1	0	38540	5	4	0	1
		<i>Podarcis muralis</i>	1.8	105207	34	34	0	0	24953	32	29	2	1	33345	0			
		<i>Zootoca vivipara</i>	1.4	59550	20	18	2	0	23328	31	30	1	0	20955	0			
	Teiidae	<i>Aspidoscelis marmoratus</i>	1.9	0	0				1575	1	1	0	0	91195	784	744	39	1
		<i>Aspidoscelis tigris steynegeri</i>	1.5	0	0				2367	2	2	0	0	94793	2152	1743	335	74
<i>Salvator merianae</i>		2.0	0	0				5467	8	7	1	0	7605	0				
Gymnophthalmidae	<i>Calyptommatius sinebrachiatus</i>	1.8	0	0				117	0				31012	36	31	3	2	
	<i>Tretosincus oriximinensis</i>	1.6	0	0				285	0				46731	609	538	62	9	
Serpentes	Typhlopidae	<i>Anilius bituberculatus</i>	1.7	200	0				1351	0				22893	0			
		<i>Pantherophis guttatus</i>	1.8	26659	2	2	0	0	497	0				10707	1238	730	387	121
	Colubridae	<i>Pantherophis obsoletus</i>	1.8	26822	1	1	0	0	578	0				10853	1948	1064	648	236
		<i>Ptyas mucosa</i>	2.4	27363	1	1	0	0	561	0				10683	1819	1088	585	146
		<i>Thermophilus baileyi</i>	1.8	10059	1	0	1	0	728	0				12872	3017	1714	1037	266
		<i>Diadophis punctatus</i>	1.8	8130	3	1	1	1	779	0				10538	2708	1031	828	849
		<i>Arizona elegans</i>	1.8	36233	5	4	0	1	958	0				15002	4252	1473	1302	1477
		<i>Thamnophis sirtalis</i>	1.9	25179	9	9	0	0	279	0				6590	164	129	26	9
		<i>Thamnophis elegans</i>	2.7	42322	10	9	1	0	668	0				8934	2203	770	679	754
		<i>Erythrocephalus jimae</i>	2.1	10123	1	1	0	0	608	0				9487	2060	1076	624	360
		<i>Hydrophis cyanocinctus</i>	2.0	6206	0				576	0				8877	1175	804	284	87
		<i>Hydrophis hartwickii</i>	2.0	5346	1	1	0	0	550	0				8423	1043	751	251	41
	<i>Hydrophis melanocephalus</i>	2.0	9216	1	1	0	0	356	0				8943	473	362	99	12	
	Elapidae	<i>Laticauda colubrina</i>	3.6	10414	0				175	0				9241	63	60	3	0
		<i>Laticauda laticaudata</i>	3.6	10377	1	1	0	0	575	0				9722	1339	927	345	67
<i>Naja naja</i>		1.5	11482	2	1	1	0	418	0				10221	1297	608	452	237	
<i>Notechis scutatus</i>		3.6	11449	0				427	0				10211	1224	803	328	93	
<i>Ophiophagus hannah</i>		1.5	9236	3	3	0	0	518	0				10561	1025	709	278	38	
<i>Pseudonaja textilis</i>		2.0	11169	2	1	1	0	505	0				10070	1773	1053	523	197	
<i>Homalopsisidae</i>		<i>Myanopsis thanlyinensis</i>	1.4	8630	5	4	1	0	668	0				9697	2643	1139	933	571
Pythonidae	<i>Python bivittatus</i>	1.4	29	0				3257	0				10775	15	11	2	2	
Boidae	<i>Charina bottae</i>	1.8	77417	23	20	3	0	3114	1	1	0	0	30677	6159	3739	1571	849	
	<i>Boa constrictor</i>	2.3	11430	0				2733	0				17746	343	287	50	6	
Viperidae	<i>Crotalus horridus</i>	1.7	3560	0				473	0				10925	929	620	261	48	
	<i>Crotalus pyrrhus</i>	1.5	2883	0				765	0				9073	1149	737	332	80	
	<i>Crotalus viridis viridis</i>	1.4	3296	0				628	0				10197	1718	871	604	243	
	<i>Protobothrops flavoviridis</i>	2.6	3732	1	1	0	0	762	0				11761	2491	1290	876	325	
	<i>Protobothrops mucrosquamatus</i>	2.4	3953	2	2	0	0	619	0				12575	3269	1448	1204	617	
	<i>Vipera berus berus</i>	2.4	36109	1	1	0	0	510	0				10041	1766	992	587	187	
Shinisauria	Shinisauridae	<i>Shinisaurus crocodilurus</i>	2.0	0	0			45589	37	37	0	0	101560	26	24	2	0	
Anguimorpha	Anguidae	<i>Elgaria multicarinata</i>	1.8	0	0			27378	20	15	3	2	35527	7	4	3	0	
		<i>Dopasia gracilis</i>	1.7	0	0			24873	5	4	1	0	28741	8	8	0	0	
Varanoidea	Varanidae	<i>Varanus komodoensis</i>	1.9	6737	0			13114	0				51059	3	3	0	0	
		<i>Phrynosoma platyrhinos</i>	1.9	16510	2	2	0	0	5799	1	0	1	0	103338	814	697	110	7
Iguania	Phrynosomatidae	<i>Sceloporus occidentalis</i>	2.9	7825	0			5329	0				93493	371	298	44	29	
		<i>Sceloporus fristicus</i>	2.5	7758	1	1	0	0	4991	0				93321	321	292	25	4
	Agamidae	<i>Calotes versicolor</i>	1.7	13638	0			3787	0				1280	0				
		<i>Pogona vitticeps</i>	1.8	35454	4	4	0	0	6143	0				5668	0			
Dactyloidea	<i>Anolis carolinensis</i>	2.2	73384	633	364	198	71	1502	0				1253	0				
Rhynchocephalia	<i>Sphenodon punctatus</i>	5.0	0	0				0	0				9362	0				

**Fig. 2** Numbers of SINE-derived satellites and the corresponding SINE loci in squamate genomes. The total number as well as that for loci with 4–5, 6–8, and more monomers are given. Relative abundance is visualized by color gradients. The only non-squamate species where Squam SINEs were found (tuatara *Sphenodon punctatus*) is given for reference. Genome sizes are given to the right of species names

*bottae* had the most sSat3 loci unearthed in this study (Fig. 2).

**Microsatellites** The  $(TC)_n$  region located within the Squam3 body (beginning at position 110 in Squam3C) is identified as a microsatellite tandem repeat site, and we observed significant variation in the length of this region considerable in some SINE copies. This region concludes prior to the end of the sSat3 tandem repeat unit; accordingly, no such dinucleotide variation was observed within sSat3. We note that the tails of most SINEs of the considered families have simple repeat sequences such as  $A_n$ ,  $(AC_2T_3)_n$ ,  $(TA_3)_n$ , or  $(CT_2)_n$ . Such regions could be deemed as SINE-derived microsatellites; however, they were neither abundant nor widespread in squamates. Therefore, we restricted our investigation to longer sSats.

#### **Squam3-derived satellites in geckos (*Gekkota*)**

Four full-length geckos' genomes were analyzed revealing that the identified sSat3s exhibited significantly greater structural variation compared to snakes.

The Squam3-derived satellite discovered in the Schlegel's Japanese gecko (*Gekko japonicus*) proved to be one of the most abundant sSat3 (and sSat) in squamates (Fig. 2). SSat3<sub>Gja</sub> can be partitioned into three variants, each with a tandem repeat unit corresponding to different sections of Squam3 (Fig. 1). The major repeat unit of sSat3<sub>Gja1</sub> corresponds to the larger central part and the 3'-terminal part of Squam3, comprising about 130 nt. Conversely, the repeat unit of sSat3<sub>Gja2</sub> corresponds to the other (5') part of Squam3 but also includes an extra ~30 nt of unknown origin, totaling approximately 150 nt. The third minor sSat3<sub>Gja3</sub> corresponds to the central part of Squam3 that covers most of the SINE head and half of LDR, making up 126 nt in length (Supplementary Fig. 4). The mean similarities of the repeat units are 86, 77, and 85%, respectively.

Furthermore, sSat3<sub>Gja1</sub> can be categorized into three variants identifiable by 16- or 23-nt truncations at the 5'-end of the tandem repeat unit (Fig. S4). The respective proportions of these variants constitute 27, 20, and 19% of all sSat3<sub>Gja</sub> loci; and the mean repeat unit similarities are 92, 85, and 87%, respectively. Noteworthy, the proportion between these variants was comparable in relatively short loci (four monomers); however, the proportion of the shortest sSat3<sub>Gja1a</sub> increased with the number of monomers and became higher in loci with more than 10 monomers (Supplementary Fig. 5).

The ocelot gecko *Paroedura picta* possesses two Squam3-derived satellite variants. The major sSat3<sub>Ppi1</sub> (92% of all sSat3<sub>Ppi</sub> loci) resembled sSat3<sub>Gja1</sub> with the exception of an additional 14-nt region of unknown origin at the repeat unit's 5' end (Fig. 1). On average, the

tandem length was 143 nt. Despite sharing a similar structure, sSat3<sub>Gja</sub> and sSat3<sub>Ppi</sub> exhibit sequence differences, as depicted in Supplementary Fig. 4. The minor variant sSat3<sub>Ppi2</sub> (6%) had a repeat unit that matched the tRNA portion of Squam3. It is predominantly present in short satellite loci and is exceedingly rare in longer ones. The average length of the repeat unit was 151 nt, while their similarities were 82 and 63%, respectively.

In the Townsend's least gecko *Sphaerodactylus townsendi*, there is a noticeable decrease in the number of Squam3-derived satellites (Fig. 2) with highly diverse repeat units that preclude grouping. However, it is evident that either 5' - or 3' -terminal regions initiated this sSat3.

In the leopard gecko *Eublepharis macularius*, Squam3-derived satellite loci were rare and highly diverse indicating unsuccessful sSat3 germination.

#### **Among the three teiid genomes**

Analyzed, sSat3 was found in both the western *Aspidoscelis tigris* and marbled *A. marmoratus* whiptails. The *A. tigris* genome contained two variants of the repeat unit, sSat3<sub>Ati1</sub> had a match with the very end of the head, the entire Squam3 body, LDR, and an additional 44 nt of unknown origin at the 3' end (215 nt), while sSat3<sub>Ati2</sub> had a slightly longer Squam3 head-related part, the entire body, and LDR (179 nt) (Supplementary Fig. 4). SSat3<sub>Ati1</sub> was ~three times more abundant than sSat3<sub>Ati2</sub> (48 and 29%) and this ratio remained stable in sSat3<sub>Ati</sub> loci of varying lengths. In *A. marmoratus*, sSat3 corresponded to sSat3<sub>Ati1</sub> but its abundance was roughly three times less despite similar numbers of Squam3 SINE (Fig. 2). The mean similarity of tandem repeat units was 97 and 84%, respectively. Both whiptail genomes showed a prevalence of relatively short tandem repeats (4–5 monomers), although the sequence similarity was higher in longer loci. SSat3 was not detected in the black and white tegu *Salvator merianae*, which matched a ~12-fold difference in the Squam3 abundance between these genera (Fig. 2).

#### **Gymnophthalmid lizards**

Exhibited relatively low numbers of sSat3 loci in the two genomes analyzed. The Oriximina lizard *Tretioscincus oriximinensis* harbored the major sSat3 variant (94% of all sSat3<sub>Tor</sub> loci), which covered most of the Squam3 central region (Fig. 1). The mean similarity of sSat3<sub>Tor</sub> repeat units was 92%. Minor variants covered most of the Squam3 body but their limits differed from one locus to another. This was also applicable to the gymnophthalmid *Calypotommatus sinebrachiatus*, which had a very small number of sSat3 loci (Fig. 2) with variable repeats.

### Phrynosomatid lizards

Had a relatively low number of sSat3 satellite loci despite numerous Squam3 SINEs. All sSat3 loci within the three analyzed genomes were short and heterogeneous (Fig. 2) with a notable exception of a distinct locus found in the western fence lizard *Sceloporus occidentalis* including as much as over 1200 relatively homogeneous monomers.

### Squam2-derived satellites

(sSat2) were largely found in geckos, although minor number of loci are scattered among lizards (Lacertidae, Shinisauridae, Anguinae, and Teiidae).

### Squam2-derived satellites in geckos

The most prevalent sSat2 was identified in *Gekko japonicus* (Fig. 2). The major sSat2<sub>Gja1</sub> (79% of all sSat2<sub>Gja</sub> loci) has the tandem repeat unit (139 nt) corresponding to the 3' part of Squam2 covering the head and one-third of the body. The second variant sSat2<sub>Gja2</sub> (14%) covers the same region truncated at both sides (115 nt) (Fig. 1 and Supplementary Fig. 6). The positions of tandem repeat units along the Squam2 sequence are presented in Supplementary Fig. 7. Short loci (4–6 middle monomers) with sSat2<sub>Gja1</sub> were roughly six times more common than those with sSat2<sub>Gja2</sub>, but the proportion of sSat2<sub>Gja2</sub> increased with the locus length. Most (but not all) sSat2<sub>Gja</sub> loci were composed of a single monomer type. The mean similarity of sSat2<sub>Gja1</sub> and sSat2<sub>Gja2</sub> repeat units was 83 and 88%, respectively; however, similarity could reach 100% in the longest loci (20–60 monomers).

The *Paroedura picta* genome does not have numerous sSat2 loci (Fig. 2), most of which are short and heterogeneous. No definite variants could be identified.

### Two other geckonid lizards

*Eublepharis macularius* and *Sphaerodactylus townsendi*, have a very low number of sSat2 loci (Fig. 2). Single relatively long loci with more than 10 monomers display uniform middle repeat units; conversely, short loci demonstrate significant heterogeneity.

### Other squamates

(lacertids, *Shinisaurus crocodilurus*, and *Elgaria multicarinata*) where sSat2 could be found also had low sSat2 numbers. The number of monomers rarely exceeded four, and no structural patterns could be revealed.

### Squam1-derived satellites

(sSat1) could be identified in the fewest number of squamate taxa. Specifically, sSat1 was only abundant in the green anole *Anolis carolinensis* (Fig. 2). The major sSat1<sub>Aca</sub> (88% of all loci) corresponds to around half of the Squam1 body (Fig. 1) but includes minor variants

with relatively long central or terminal deletions. The sSat1<sub>Aca</sub> repeat unit is 145 nt long (Supplementary Fig. 9) and its mean similarity is 81%. The minor variants largely confined to regions nearer to the Squam1 head.

In other taxa, sSat1 was found in low numbers of loci without a consistent pattern, although Squam1 SINE was present in high abundance in the genomes of lacertids and some other squamates (Fig. 2).

### Variation of orthologous sSat3 loci

We examined the variation of SINE-derived satellites using sSat3<sub>SRP</sub> as an example in two rattlesnakes, *Crotalus pyrrhus* and *C. viridis*. We identified all pairs that contained a minimum of two sSat3<sub>SRP</sub> repeat units in one genome and had matching flanking sequences in the other genome. Most loci had the same number of tandem repeat units (74%) or differed by one (14%). However, some loci varied by 10 or more repeat units (Supplementary Fig. 8A). Examination of these orthologous pairs revealed significant variation in their structure beyond the number of tandems. Some pairs consisted of only tandem fragments or included insertions of unknown origin. Many also contained stretches of N resulting from poor sequencing. Three examples are shown in Supplementary Fig. 8B. The top one represents the majority of identified loci with the same number of repeat units. The middle diagram depicts a relatively rare example with distinct tandem numbers (note the shortened rightmost repeat unit in both genomes). The bottom diagram shows an uncommon variant with both Squam3 SINE termini and several tandem repeats in between. This pilot analysis showcased a promising approach and should be expanded to incorporate additional genomes and a wider variety of species. However, further research is required to explore this matter fully.

### Chromosomal localization of SINE-derived satellites

Satellites are a component of heterochromatin, specifically, the pericentromeric and subtelomeric domains [1], in contrast to microsatellites and retrotransposons, which are mostly found in euchromatin. We investigated the chromosomal localization of sSat3, the most consistent of the SINE-derived satellites, in three snake genomes assembled to near-chromosome level (Indian cobra *Naja naja*, Lataste's viper *Vipera latastei*, and many-banded krait *Bungarus multicinctus*). The assemblies are fairly accurate but not flawless, particularly in the problematic large repetitive regions. The assemblies comprise some fragments ("chromosomes") that are relatively long ( $\geq 100$  Mbp) and a plethora of shorter ones.

Typically, sSat3 appeared in the subtelomeric areas in the best assembled chromosomes (not necessarily the largest); other fragments could have these repeats at a

single end of the fragment. Most of the shorter fragments only displayed sporadic occasions of sSat3.

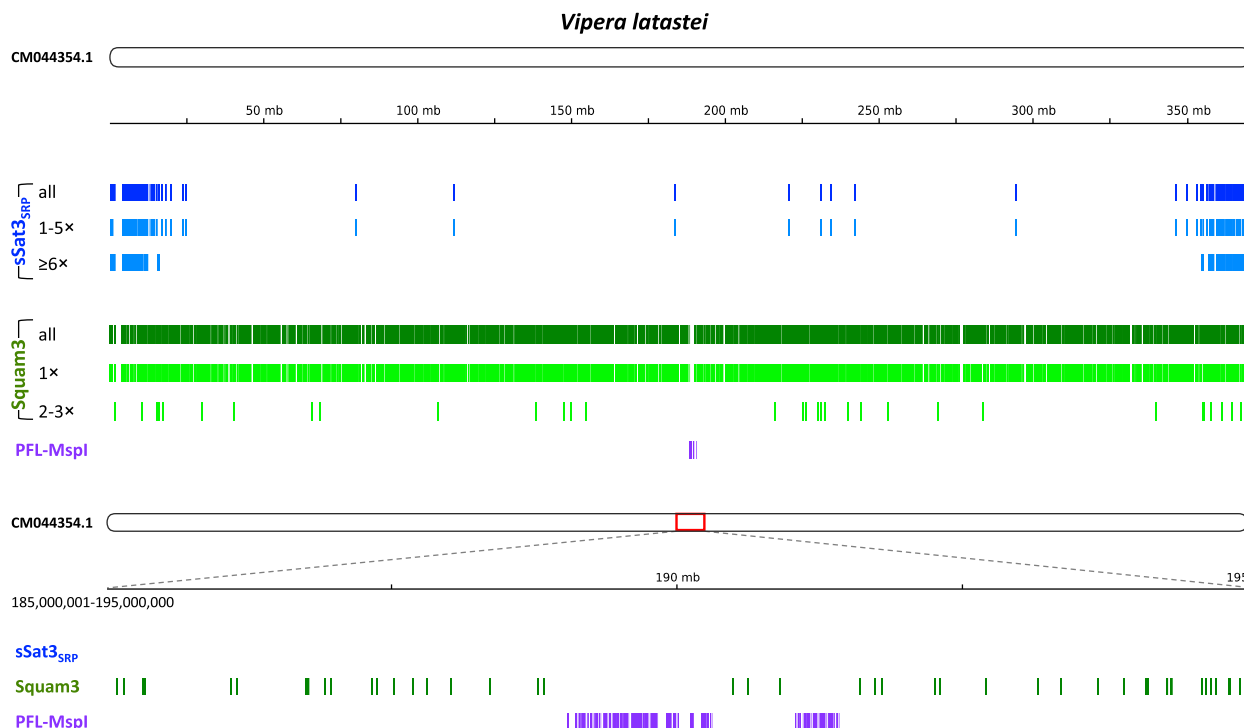
There is only one fragment/chromosome (in *N. naja*) that may indicate a pericentromeric localization of such repeats (Supplementary Fig. 10B). To our knowledge, only three centromeric repeats were reported in two snake species [17]. Our search for these repeats revealed only a few (0–10) such sequences in the genomes of Indian cobra and many-banded krait. However, one of them (PFL-MspI) from *Protobothrops flavoviridis* (Viperidae) proved abundant (>1000 of loci ~15% of which included more than a hundred tandems) in the genome of the Lataste’s viper; their consensus sequence proved 68% identical to PFL-MspI (Supplementary Fig. 10A). Moreover, they localized to two neighboring regions in an internal part of all properly assembled chromosomes and were arranged in long tandem arrays (Fig. 3), suggesting that this PFL-MspI is a pericentromeric satellite in the Lataste’s viper as well.

We evaluated the distribution of longer and shorter stretches of tandem repeats and found a scarcity of long stretches (more than 10×) within the internal chromosome regions. This could be a result of fewer number of longer sequences. Squam3 sequences were used as a control, displaying an expected random distribution.

Recently, a highly accurate squamate genome assembly was published for the leopard gecko *Eublepharis macularius*. Telomeric termini were identifiable on certain chromosomes (Supplementary Fig. 10C). While sSat3 is less abundant than in most snakes, there is tendency to subtelomeric regions. The distribution of rare sSat2 is indeterminate.

We used the top-rated vertebrate genomic assembly (human T2T-CHM13v2.0) as a reference for visualizing the distribution of selected satellites and retrotransposons (Supplementary Fig. 10D). Generally, satellites largely localized in the (peri)centromeric regions. Specifically, α-satellites mark centromeres, while pericentromeric regions are occupied by HSATII satellites and/or (GGAAT)<sub>n</sub> microsatellites, particularly in chromosomes with very long pericentromeric regions (e.g., 1, 9, and 16). The predominant SINES and LINES (Alu and L1) are evenly distributed throughout chromosomes but are absent from the centromeric and long pericentromeric regions.

The pattern observed for Squam3 SINE and PFL-MspI satellite was mostly replicated (Fig. 3 and Supplementary Fig. 10). However, sSat3 unequivocally exhibited subtelomeric rather than pericentromeric localization in the three snake genomes. Additionally, the CENP-B motif,



**Fig. 3** Chromosomal localization of sSat3 (blue), Squam3 SINE (green), and putative centromeric satellite PFL-MspI (purple) in the largest chromosome assembly (CM044354.1, 370 Mbp) of the Lataste’s viper *Vipera latastei*. Size fractions corresponding to 1–5 and ≥6 tandems for sSat3 as well as 1 and 2–3 Squam3 tandem repeats are shown below in lighter shades. The lower panel is the putative pericentromeric area (10 Mbp, 185,000,001–195,000,000) containing PFL-MspI and Squam3 but no sSat3



a marker for specific centromeric satellites [18], was not detected.

## Discussion

SINE-derived satellites were identified based on their similarity with SINEs. In squamates, most sSats are tandem repeats with a longer terminal unit on either end (3' or 5') sharing similarity with the corresponding SINE region up to its terminus (Fig. 1). This pattern was common for all major sSat instances identified in this study. However, there were minor variants without such extended terminal units or with two of them corresponding to both SINE termini (as shown in the lower panel of Supplementary Fig. 8B).

Some genomes have a small number of sSat loci with only a few tandem repeat units resulting in significant variation in the corresponding SINE regions across loci (as seen in *Python bivittatus*, Supplementary Fig. 2). In contrast, other genomes have multiple and lengthy sSat loci, containing up to 1000 repeat units. These sSat loci display high homogeneity of tandem sequences both within and between loci, and this homogeneity increases with the number of repeat units per locus (Supplementary Fig. 5). This is in line with the concept of concerted evolution of satellites [1], indicating that only the abundant sSat variants undergo homogenization. However, even the genomes rich in sSats contain a few short loci with variable repeat units. Please note that most, if not all, of these numeric values are underestimated due to poor genome sequencing (which is especially conspicuous in repetitive DNA regions).

Satellites discovered come from three SINE families, namely, Sauria (Squam1), Squam2, and Squam3. Squam3 are more or less evenly distributed throughout squamates (and are even found in the tuatara of the distinct order Rhynchocephalia). Squam2 SINEs, on the other hand, are present in all squamates with varying genomic copy numbers that can differ by up to 500-fold in extreme cases. Squam1 presents a random distribution in squamates and is absent in approximately 15% of genomes tested (Fig. 2). This distribution pattern was somewhat replicated by sSats. Notably, no sSat1 was detected in the genomes lacking Squam1, including teiid (Teiidae) and spectacled (Gymnophthalmidae) lizards as well as species with a low number of SINE copies numbered in the hundreds, such as Squam2 in most snakes. Nevertheless, there are numerous genomes with a relatively high number of SINE copies that lack derived satellites, for example, Squam3 in lacertids. Likewise, no significant correlation was found between the abundance of sSats (as well as SINEs) and the genome size (Fig. 2).

The taxonomic variation of sSats was irregular. Excluding two exceptions being snakes presenting amazingly

uniform sSat<sub>SRP</sub> and two *Aspidoscelis* whiptails, most distant taxa possessed differing sSat repeat unit structures (Fig. 1). This pattern is typical for species-specific satellites but can also be observed in taxa of different ranks [1]. The number of sSat loci exhibits significant variation and can exceed 20 fold in representatives of the same genus (*Laticauda laticaudata* and *Laticauda colubrina*). This type of variation is not uncommon for satellites [2].

It should be noted that proper SINEs can also form tandem dimers, trimers, and so on. The frequency of tandem SINEs is generally low, but in some cases, it can be significant. For example, nearly half of Dip SINE copies in the genome of Gobi jerboa *Allactaga bullata* exist in multimers consisting of full-length Dip and/or B1 SINEs with tandems reaching up to eight [19]. At the same time, there are SINEs that amplify (i.e., are transcribed and integrated into the genome) as dimeric tandems such as the well-known *Alu* [20]; dimeric SINEs are known in vertebrates, insects, and plants [21]. In addition, SINEs with short (~30nt) internal duplications were also reported [22, 23]. We found no true dimeric Squam SINEs in the genomes analyzed. The repetitive elements discussed embody a vast spectrum of structures from SINEs bearing extended internal duplications to satellite-like tandem repeats with primordial SINE sequences at either end.

The lengths of repeat units constitute a range from 150 to 180bp or 300 to 360bp in numerous satellites [2]. It is a prevalently accepted hypothesis that links these lengths to the DNA length wrapped around one or two nucleosomes [24]. It may be of interest in this context to note the alternating (~di and trimeric) pattern of monomers observed in the glossy snake *Arizona elegans* (Supplementary Fig. 3) and some other species, suggesting that the elongation of such satellites involves dimeric rather than monomeric units.

Most of sSat variants appear to have independent origins in different squamates. However, sSat3 in snakes, particularly in higher snakes, presents a contrasting opposite example (Supplementary Fig. 1). Another example, though less conclusive, is sSat3<sub>ppi</sub> in the gecko *Paroedura picta*, which only matches a subsubvariant sSat3<sub>Gja1b</sub> in *Gekko japonicus* (Supplementary Fig. 4).

Analysis of sSats in genomes with high sSats abundance indicates that initially diverse variants tend to evolve into a few or even a single variant that dominates in terms of locus number, length, and sequence homogeneity. In the case of *Gekko japonicus*, three sSat3 subvariants are equally represented in short loci while long loci are largely composed of sSat3<sub>Gja1</sub> (Supplementary Fig. 5).

Overall, certain primary duplications are more favorable as satellite targets for further amplification, while most remain as single genomic instances. The genome's

susceptibility to repeat amplification also plays an important role. For example, *Gekko japonicus* exhibits a particularly high abundance of sSat variants, and its genome is larger than that of most other squamates. However, there are other species with even larger genomes (e.g., the sea krait *Laticauda colubrina*), but they exhibit almost no sSat. The factor that makes organisms susceptible to this remains to be revealed.

According to Garrido-Ramos, different families of satellites tend to occupy the same regions of chromosomes, where heterochromatin amplification is initiated [1], mostly pericentromeric or subtelomeric regions. Our findings indicate that sSats have a clear preference for subtelomeric regions, which is more pronounced for longer loci (Fig. 3 and Supplementary Fig. 10), though they are not as abundant.

The concept of satellite emergence [2, 25] is a consistent process with three distinct steps. First, individual genomic sequences duplicate to create a “library” of short tandem repeats. Second, some of these repeats undergo further amplification, the mode of which can differ in descendant species. Finally, concerted evolution homogenizes the most successful satellite variants.

The data from this study of SINE-derived satellites in squamates generally support this concept with some specific variations. The large number of genomic SINE copies significantly enhances the likelihood of initial duplication via unequal homologous recombination [26]. Additionally, insertion into the same locus becomes much more probable.

The SINE-derived satellites identified in squamates have only reached the stage of approaching concerted evolution and are not as homogeneous as other satellites. However, the mechanisms of homogenization are more efficient for longer tandem arrays [27]; likewise, we have observed that long sSat<sub>SRP</sub> loci contain more uniform repeat sequences.

We did not come across such satellites in our studies of mammalian SINEs, although we haven't specifically checked. Therefore, it is uncertain whether these SINE-derived satellites are restricted to squamates. Nonetheless, this approach appears to have the potential to investigate the satellite evolution and possibly broader aspects of genome existence.

## Methods

Genomic data were downloaded from NCBI Genomes (<https://www.ncbi.nlm.nih.gov/genome>). The following assemblies were used: Schlegel's Japanese gecko *Gekko japonicus*, Gekko\_japonicus\_V1.1; ocelot gecko *Paroedura picta*, Ppicta\_v2.0; Townsend's least gecko *Sphaerodactylus townsendi*, MPM\_Stown\_v2.3; Valentin's lizard

*Darevskia valentini*, Dval\_245; sand lizard *Lacerta agilis*, rLacAgi1.pri; European green lizard *Lacerta viridis*, ASM90024590v1; western green lizard *Lacerta bilineata*, *L. bilineata* genome assembly; common wall lizard *Podarcis muralis*, PodMur\_1.0; viviparous lizard *Zootoca vivipara*, UG\_Zviv\_1; marbled whiptail *Aspidoscelis marmoratus*, AspMar1.0; western whiptail *Aspidoscelis tigris stejnegeri* rAspTig1.0.p; Argentine black and white tegu *Salvator merianae* HLtupMer6; lesser microteiid *Calyptommatus sinebrachiatus*, HLcalSin1; Oriximina lizard *Tretioscincus oriximinensis*, HLtreOri1; prong-snouted blind snake *Anilius bituberculatus*, ASM2237905v1; corn snake *Pantherophis guttatus*, UNIGE\_PanGut\_3.0; western rat snake *Pantherophis obsoletus*, UNIGE\_PanObs\_1.0; Dhaman *Ptyas mucosa*, UNIGE\_Pmuc\_v1.0; Bailey's snake *Thermophis baileyi*, DSBC\_Tbai\_1.0; ringneck snake *Diadophis punctatus similis*, rDiaPun1.0; glossy snake *Arizona elegans*, rAriEle1.0.p; eastern garter snake *Thamnophis sirtalis*, *Thamnophis sirtalis*-6.0; western terrestrial garter snake *Thamnophis elegans*, rThaEle1.pri; turtlehead sea snake *Emydocephalus ijimae*, emylji\_1.0; Asian annulated sea snake *Hydrophis cyanocinctus*, ASM402372v1; Hardwick's sea snake *Hydrophis hardwickii*, ASM402376v1; slender-necked sea snake *Hydrophis melanocephalus*, hydMel\_1.0; yellow-lipped sea krait *Laticauda colubrina*, latCor\_2.0; blue-ringed sea krait *Laticauda laticaudata*, latLat\_1.0; Indian cobra *Naja naja*, NN\_10x\_BNG; mainland tiger snake *Notechis scutatus*, TS10Xv2-PRI; king cobra *Ophiophagus hannah*, OphHan1.0; eastern brown snake *Pseudonaja textilis*, EBS10Xv2-PRI; *Myanophis thanlyinensis*, ASM1765603v1; Burmese python *Python bivittatus*, Python\_molurus\_bivittatus-5.0.2; northern rubber boa *Charina bottae*, rChaBot1.0.p; timber rattlesnake *Crotalus horridus*, ASM162548v1; southwestern speckled rattlesnake *Crotalus pyrrhus*, CroTMitch1.0; prairie rattlesnake *Crotalus viridis viridis*, UTA\_CroVir\_3.0; habu *Protobothrops flavoviridis*, HabAm\_1.0; brown spotted pitviper *Protobothrops mucrosquamatus*, P.Mucros\_1.0; common viper *Vipera berus berus*, Vber.be\_1.0; Chinese crocodile lizard *Shinisaurus crocodilurus*, IOZ\_Scro\_1.0; southern alligator lizard *Elgaria multicarinata webbia*, rElgMul1.0.p; Komodo dragon *Varanus komodoensis*, ASM479886v1; desert horned lizard *Phrynosoma platyrhinos*, MUOH\_PhPlat\_1.1; western fence lizard *Sceloporus occidentalis*, rSceOcc1.0.p; plateau fence lizard *Sceloporus tristichus*, ASM1680106v1; garden lizard *Calotes versicolor*, ASM2071127v1; central bearded dragon *Pogona vitticeps*, pvi1.1; and green anole *Anolis carolinensis*, AnoCar2.0. The genomes of the boa *Boa constrictor*, leopard gecko *Eublepharis macularius*, and Asian glass lizard *Dopasia gracilis* were obtained from various sources.

To refine the consensus SINE sequences in the studied genomes, we searched for close to full-length SINE copies with more than 90% length and more than 65% similarity to the consensus using the ssearch36 program from the FASTA package, followed by analysis of at least 100 randomly selected copies, and using the SubFam script. Briefly, the script partitioned a set of sequences of individual SINE loci, presorted by similarity using MAFFT [28] into groups of 100 copies, created a multiple alignment and a consensus sequence for each group, and all such consensus sequences were in turn aligned. The resulting alignment was used to estimate the diversity of SINEs. In each genome, all sequences with similarities of at least 20% of the consensus sequence length were searched for the three families (Squam1, Squam2, and Squam3). This size of the SINE fragments found allowed us to detect all genomic loci potentially containing tandemly duplicated SINE fragments. The genomic coordinates of fragments found less than 100 bp apart were merged together, and those that were longer than 500 bp after fusion were considered as candidates for containing a satellite.

Each candidate locus was analyzed using custom scripts to reveal tandemly repeated SINE-derived regions. The loci containing more than four such tandems within 100 bp from each other were considered as an sSat. The sSat monomers were counted, aligned, and comparatively analyzed between and within the loci. This made it possible to reveal tandem subvariants and to evaluate sSat diversity including the relative frequency of sSat variants in individual loci and in the whole genome.

For several snake species with relatively good genome assemblies, the chromosomal localization of the found sSat loci was analyzed in comparison with certain tandem and interspersed repeats using the Integrative Genomics Viewer (<https://igv.org>; IGV-Web app v.1.12.5).

Amazingly, overlapping tandem repeats such as (ATGGA)<sub>n</sub>, (TGGAA)<sub>n</sub>, etc. and their complementary variants are annotated separately by RepeatMasker; this also applies to their ~dimer (ATcGAATGGA)<sub>n</sub> etc. All these were combined as (GGAAT)<sub>n</sub> in Supplementary Fig. 10.

Orthologous sSat3<sub>SRP</sub> loci in the genomes of the southwestern speckled rattlesnake *Crotalus pyrrhus* and prairie rattlesnake *Crotalus viridis* were analyzed by domestic bash and Perl scripts using the following pipeline. Briefly, (1) identification of sSat3<sub>SRP</sub> tandem repeats (starting from two) in one genome; (2) identification of loci in the second genome with both flanking regions at a reasonable distance from each other; (3) analysis of found pairs in two genomes without flanking sequences.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13100-023-00309-2>.

**Additional file 1** Multiple alignment of consensus sequences of sSat3 repeat units in snakes. The Squam3 consensus sequence is given above. Subscript indices indicate the following species. Pgu, *Pantherophis guttatus*; Lco, *Laticauda colubrina*; Pob, *Pantherophis obsoletus*; Ael, *Arizona elegans*; Pte, *Pseudonaja textilis*; Cho, *Crotalus horridus*; Cvi, *Crotalus viridis viridis*; Pfl, *Protobothrops flavoviridis*; Prm, *Protobothrops mucrosquamatus*; Vbe, *Vipera berus*; Pmuc, *Ptyas mucosa*; Dpu, *Diadophis punctatus*; Tel, *Thamnophis elegans*; Tsi, *Thamnophis sirtalis*; Hcy, *Hydrophis cyanocinctus*; Hha, *Hydrophis hardwickii*; Hme, *Hydrophis melanocephalus*; Nna, *Naja naja*; Oha, *Ophiophagus hannah*; Mth, *Myanophis thanlyinensis*; Tba, *Thermophilis baileyi*; Eij, *Emydocephalus ijimae*; Lla, *Laticauda laticaudata*; Nsc, *Notechis scutatus*; Cpy, *Crotalus pyrrhus*; Bco1 and Bco2, *Boa constrictor*; and Cbo, *Charina bottae*. The snake families are colored: Colubridae, yellow; Elapidae, green; Viperidae, blue; Boidae, red; and Homalopsidae, magenta.

**Additional file 2.** A. Multiple sequence alignment of consensus sequences of repeat units of Squam3-derived satellites from the Burmese python *Python bivittatus*. While all these sequences share similarity with the Squam3 consensus (upper) sequence, they correspond to different SINE regions and have little in common with each other. B. Three examples of sSat3<sub>Pbi</sub> loci. The upper one (Pbi-672 in panel A) has the repeat unit covering the body of Squam3; the leading and trailing monomers also contain the 5' and 3' parts of Squam3, respectively. The repeat unit of the second one (Pbi-809) corresponds partially to the Squam3 head and body but has an extra 62-nt sequence of unknown origin at the 3' end. The lower one (Pbi-558) demonstrates a highly irregular structure.

**Additional file 3.** Three examples of alternating tandem repeat units in Squam3-derived satellites in the glossy snake *Arizona elegans*. Each multiple alignment includes consecutive middle repeat units of a particular genomic locus (specified above) and the top consensus sequence. Background sequence colors mark different variants of tandem repeat units.

**Additional file 4.** Multiple alignment of consensus sequences of sSat3 in geckos, teiids, and gymnophthalmids. Names on the left indicate corresponding species. Gja, *Gekko japonicus*; Ppi, *Paroedura picta*; Ati, *Aspidoscelis tigris*; Ama, *Aspidoscelis marmoratus*; Tor, *Tretioscincus oriximinensis*. Indices after the species code indicate (sub)variants. The lizard families are colored: Gekkota, yellow; Teiidae, green; and Gymnophthalmidae, blue. Squam3 SINE consensus sequence specific for geckos and lacertids is given above.

**Additional file 5.** Distribution of three subvariants of sSat3<sub>Gja</sub> as a function of number of repeat units in the loci of the Schlegel's Japanese gecko *Gekko japonicus*. The ordinate represents the number of repeat units in loci. The abscissa shows the proportion of each sSat3<sub>Gja</sub> subvariant (A) or their absolute number (B).

**Additional file 6.** Multiple alignment of consensus sequences of sSat2 in the Schlegel's Japanese gecko *Gekko japonicus*. The gecko-specific Squam2 consensus sequence is given above.

**Additional file 7.** Distribution of nucleotide positions of sSat2<sub>Gja</sub> with 10 middle monomers along Squam2 consensus sequence. Each position of all monomers was compared against Squam2 sequence and assigned a weight of 10 for a match; 1, for mismatch, and 0 for a gap. The cumulative statistics is presented. Abscissa: Squam2 consensus position; ordinate: total position weight.

**Additional file 8.** Variation of orthologous sSat3<sub>SRP</sub> loci in two snakes, southwestern speckled rattlesnake *Crotalus pyrrhus* and prairie rattlesnake *Crotalus viridis*. A. Distribution of orthologous sSat3<sub>SRP</sub> loci by difference in the number of tandem repeat units. B. Examples of orthologous sSat3<sub>SRP</sub> loci in *C. pyrrhus* and *C. viridis* genomes. The red arrows indicate the repeat units, while the green and blue ones correspond to the 5' and 3' ends of Squam3 SINE. No flanking genomic sequences are shown. Genomic affiliations are given above diagrams.

**Additional file 9.** Sequence alignment of Squam1 SINE and sSat1<sub>Aca</sub> repeat unit from the green anole *Anolis carolinensis*.

**Additional file 10.** Chromosomal localization of certain genomic repeats in selected chromosomes of some squamates and human (for reference). A. Chromosomal localization of certain genomic repeats in selected chromosomes of many-banded krait *Bungarus multicinctus* and Indian cobra *Naja naja*. Blue bars correspond to sSat3, light blue bars show stretches of 1–5 and  $\geq 6$  tandems; and green bars show Squam3 SINE. B. Chromosomal localization of certain genomic repeats in selected chromosomes of leopard gecko *Eublepharis macularius*. Dark and light blue bars correspond to sSat3 and sSat2, respectively; dark and light green bars correspond to Squam3 and Squam2, respectively; T<sub>2</sub>AG<sub>3</sub> are stretches of at least 1000 repeats marking telomeres. C. Chromosomal localization of certain genomic repeats in the human genome (T2T-CHM13): euchromatic interspersed retrotransposons L1 LINE (brown) and Alu SINE (green) as well as heterochromatic satellites (blue) including centromeric  $\alpha$ -satellite (light blue) and pericentromeric HSAT2 and GGATT microsatellites (purple) as annotated by Repeat Masker (see Materials and Methods). The upper panel shows the largest chromosome 1 (248 Mbp) with an extended pericentromeric region, while the lower panel represents a typical chromosome 7 (161 Mbp).

#### Authors' contributions

NSV and SAK, conceptualization; all, genomic data analysis; NSV and SAK, study design and manuscript preparation; APR, supervision; APR, project administration and funding acquisition. All authors read and approved the final manuscript. Sergei Kosushkin and Nikita Vassetzky contributed equally to this work.

#### Funding

This research was funded by the Russian Science Foundation (RSF) Project No. 19–14–00083.

#### Availability of data and materials

The data generated are available in the manuscript supporting files.

#### Declarations

##### Ethics approval and consent to participate

Not applicable.

##### Consent for publication

Not applicable.

##### Competing interests

The authors declare no competing interests.

Received: 24 October 2023 Accepted: 24 November 2023

Published online: 07 December 2023

#### References

- Garrido-Ramos MA. Satellite DNA: an evolving topic. *Genes* (Basel). 2017;8(9):230.
- Plohl M, Meštrović N. Satellite DNA. *Evolution*. 2012;7:126–52.
- Kramerov DA, Vassetzky NS. SINEs. *Wiley Interdiscip Rev RNA*. 2011;2:772–86.
- Kramerov DA, Vassetzky NS. Origin and evolution of SINEs in eukaryotic genomes. *Heredity*. 2011;107:487–95. Available from: [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list\\_uids=21673742](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=21673742)
- Li S, Shen X. Long interspersed nuclear element 1 and B1/Alu repeats blueprint genome compartmentalization. *Curr Opin Genet Dev*. 2023;80:102049. <https://doi.org/10.1016/j.cde.2023.102049>
- Meštrović N, Mravinac B, Pavlek M, Vojvoda-Zeljko T, Šatović E, Plohl M. Structural and functional liaisons between transposable elements and satellite DNAs. *Chromosom Res*. 2015;23:583–96.
- Paço A, Freitas R, Vieira-Da-Silva A. Conversion of DNA sequences: from a transposable element to a tandem repeat or to a gene. *Genes* (Basel). 2019;10(12):1014.
- Nagahashi S, Endoh H, Suzuki Y, Okada N. Characterization of a tandemly repeated DNA sequence family originally derived by retroposition of tRNA<sup>Glu</sup> in the newt. *J Mol Biol*. 1991;222:391–404.
- Ackerman EJ. Molecular cloning and sequencing of OAX DNA: an abundant gene family transcribed and activated in *Xenopus* oocytes. *EMBO J*. 1983;2:1417–22.
- Lam BS, Carroll D. Tandemly repeated DNA sequences from *Xenopus laevis*. *J Mol Biol*. 1983;165:587–97.
- Bucci S, Raggianti M, Mancino G, Petroni G, Guerrini F, Giampaoli S. Rana/pol III: a family of SINE-like sequences in the genomes of western Palearctic water frogs. *Genome*. 1999;42:504–11.
- Batistoni R, Pesole G, Marracci S, Nardi I. A tandemly repeated DNA family originated from SINE-related elements in the European plethodontid salamanders (Amphibia, Urodela). *J Mol Evol*. 1995;40:608–15. Available from: [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list\\_uids=7643411](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=7643411)
- Vassetzky NS, Kosushkin SA, Korchagin VI, Ryskov AP. New Ther1-derived SINE Squam3 in scaled reptiles. *Mob DNA*. 2021;12:1–10.
- Piskurek O, Austin CC, Okada N. Sauria SINEs: novel short interspersed retroposable elements that are widespread in reptile genomes. *J Mol Evol*. 2006;62:630–44. Available from: <http://www.ncbi.nlm.nih.gov/entrez/pubmed/16612539>
- Kosushkin SA, Borodulina OR, Grechko VV, Kramerov DA. A new family of interspersed repeats from squamate reptiles. *Mol Biol*. 2006;40:378–82. Available from: [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list\\_uids=16637281](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=16637281)
- Grechko VV, Kosushkin S, a, Borodulina OR, Butaeva FG, Darevsky IS. Short interspersed elements (SINEs) of squamate reptiles (Squam1 and Squam2): structure and phylogenetic significance. *J Exp Zool*. 2011;316B:212–26. Available from: [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list\\_uids=21462315](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=21462315)
- Matsubara K, Uno Y, Srikulnath K, Seki R, Nishida C, Matsuda Y. Molecular cloning and characterization of satellite DNA sequences from constitutive heterochromatin of the habu snake (Protobothrops flavoviridis, Viperidae) and the Burmese python (*Python bivittatus*, Pythonidae). *Chromosoma*. 2015;124:529–39.
- Masumoto H, Nakano M, Ohzeki JI. The role of CENP-B and  $\alpha$ -satellite DNA: De novo assembly and epigenetic maintenance of human centromeres. *Chromosom Res*. 2004;12:543–56.
- Vassetzky NS, Borodulina OR, Ustyantsev IG, Kosushkin SA, Kramerov DA. Analysis of SINE families B2, dip, and ves with special reference to polyadenylation signals and transcription terminators. *Int J Mol Sci*. 2021;22(18):9897.
- Quentin Y. 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*. 1992;20:3397–401.
- Vassetzky NS, Kramerov DA. SINEBase: a database and tool for SINE analysis. *Nucleic Acids Res*. 2013;41:D83–9. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531059&tool=pmcentrez&rendertype=abstract>
- Zietkiewicz E, Labuda D. Mosaic evolution of rodent B1 elements. *J Mol Evol*. 1996;42:66–72. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/8576966>
- Vassetzky NS, Ten OA, Kramerov DA. B1 and related SINEs in mammalian genomes. *Gene*. 2003;319:149–60.
- Henikoff S, Ahmad K, Malik HS. The centromere paradox: Stable inheritance with rapidly evolving DNA. *Science*. 2001;293:1098–102.
- Ruiz-Ruano FJ, López-León MD, Cabrero J, Camacho JPM. High-throughput analysis of the satellitome illuminates satellite DNA evolution. *Sci Rep*. 2016;6:1–14. <https://doi.org/10.1038/srep28333>
- Kramerov DA, Vassetzky NS. Short retroposons in eukaryotic genomes. *Int Rev Cytol*. 2005;247:165–221. Available from: [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list\\_uids=16344113](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=16344113)
- Pavlek M, Gelfand Y, Plohl M, Meštrović N. Genome-wide analysis of tandem repeats in *Tribolium castaneum* genome reveals abundant and

highly dynamic tandem repeat families with satellite DNA features in euchromatic chromosomal arms. *DNA Res.* 2015;22:387–401.

28. Yamada KD, Tomii K, Katoh K. Application of the MAFFT sequence alignment program to large data - reexamination of the usefulness of chained guide trees. *Bioinformatics.* 2016;32:3246–51.

### **Publisher's Note**

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

