bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

On the contingent nature of satellite DNA evolution

Juan Pedro M. Camacho[1], Josefa Cabrero[1], María Dolores López-León[1], María Martín-Peciña[1], Francisco Perfectti[1], Manuel A. Garrido-Ramos[1], Francisco J. Ruiz-Ruano[2,3,*]

[1]Departamento de Genética, Universidad de Granada, 18071, Granada, Spain

[2]Department of Organismal Biology – Systematic Biology, Evolutionary Biology Centre, Uppsala University, SE-752 36, Uppsala, Sweden

[3]School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, NR4 7TU, UK

[*]Corresponding author: Francisco J. Ruiz-Ruano (francisco.ruiz-ruano@ebc.uu.se, f.ruiz-ruano-campana@uea.ac.uk)

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Background: The full catalogue of satellite DNA (satDNA) within a same genome constitutes the satellitome. The Library Hypothesis predicts that satDNA in relative species reflects that in their common ancestor, but the evolutionary mechanisms and pathways of satDNA evolution have never been analyzed for full satellitomes. We compare here the satellitomes of two Oedipodine grasshoppers (*Locusta migratoria* and *Oedaleus decorus*) which shared their most recent common ancestor about 22.8 Ma ago.

Results: We found that about one-third of their satDNA families (near 60 in every species) showed sequence homology, and were grouped into 12 orthologous superfamilies. The turnover rate of consensus sequences was extremely variable among the 20 orthologous family pairs analyzed in both species. The satDNAs shared by both species showed poor association with sequence signatures and motives frequently argued as functional, except for short inverted repeats allowing short dyad symmetries and non-B DNA conformations. Orthologous satDNAs frequently showed different FISH pattern at both intra- and interspecific levels. We defined indices of homogenization and degeneration, and quantified the level of incomplete library sorting between species.

Conclusions: Our analyses revealed that satDNA degenerates through point mutation and rejuvenates through partial turnovers caused by massive tandem duplications (the so-called satDNA amplification). Remarkably, satDNA amplification increases homogenization, at intragenomic level, and diversification between species, thus constituting the basis for concerted evolution. We suggest a model of satDNA evolution by means of recursive cycles of amplification, degeneration, and rejuvenation, leading to mostly contingent evolutionary pathways where concerted evolution emerges promptly after lineages split.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Keywords: Satellite DNA, Library Hypothesis, Satellitome Evolution, Cytogenomics.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Background

Satellite DNA (satDNA) was first described by Kit (1961) in mouse and guinea-pig DNA with its repetitive nature demonstrated by Waring and Britten (1966). The first model for satDNA evolution was devised by Smith (1976), who demonstrated that DNA sequences that are not maintained by natural selection evolve a tandem repeat structure due to unequal crossing-over. Later, theoretical analyses assumed that satDNA evolution usually depends on mutation, unequal crossing-over, and random drift, with purifying selection controlling for excessive copy number (Kimura and Ohta 1979; Ohta 1981, 1983; Ohta and Kimura 1981; Stephan 1986, 1987, 1989; Charlesworth et al. 1986).

Changes in satDNA amount are mainly due to unequal crossing-over, although other mechanisms have been proposed to explain both amplification and spread of satDNA repeats (for review, see Garrido-Ramos 2017). Walsh (1987) proposed the replication of extrachromosomal circles of tandem repeats by the rolling-circle mechanism and reinsertion of replicated arrays as a powerful satDNA amplification process, a mechanism for which Cohen et al. (2005, 2010) have found some support. Additionally, transposition may operate in satDNA emergence and amplification (Šatović and Plohl 2013; Pavlek et al. 2015; Meštrović et al. 2015; Šatović et al. 2016). Ultimately, replication-slippage might be an amplification process (Stephan 1989; Walsh 1987), mainly involved in lengthening satellite monomers from basic shorter ones (Ruiz-Ruano et al. 2018a).

To explain the conservation of satellite sequences over long evolutionary periods, Fry and Salser (1977) suggested the Library Hypothesis. According to this hypothesis, a group of related species should share a common library of satDNA sequences that mostly show quantitative differences among species due to differential amplification.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Therefore, a given member of the library may appear as an abundant satDNA, while others remain at low amounts and technically undetectable. Now we know that the former can be visualized by FISH and the latter discovered by next-generation sequencing (Ruiz-Ruano et al. 2016). Fry and Salser (1977) suggested that an essential step in the evolution of some satDNA families may be the acquisition of a biological function, in which case natural selection would conserve its sequence for long evolutionary periods (Djupedal et al. 2009; Schueler et al. 2010; Fachinetti et al. 2015).

There are some examples of satDNA persisting for long, i.e., more than 40-100 Ma (see Arnason et al. 1992; Garrido-Ramos et al. 1995, 1999; de la Herrán et al. 2001a,b; Mravinac et al. 2002, 2005; Robles et al. 2004; Cafasso and Chinali 2014). Whereas the conservation of functional satDNA repeats is explained by purifying selection (see references above), the persistence over time of other satDNA arrays lacking apparent function might be simply due to chance events (Stephan 1986, 1987, 1989; Walsh 1987; Harding et al. 1992). Therefore, whether satDNA conservation in two or more species is just chance or due to selective events remains unanswered.

Dover (1982a,b) suggested unequal crossing-over, gene conversion, and transposition as molecular drive mechanisms for the concerted fixation of paralogous variants, which operate independently of natural selection and drift. Recently, this evolutionary pattern has been replaced by the birth-and-death model in the case of coding multigene families (Nei and Rooney 2005; Eirín-López et al. 2012). Concerted evolution implies that paralogous copies are more homogenized than orthologous ones when two species are compared. SatDNA families comprise thousands or millions of copies of non-coding paralogous repeat units, frequently arranged in many short arrays spread at different genomic locations (Kuhn et al. 2012; Brajković et al. 2012; Feliciello et al. 2015; Pavlek et al. 2015; Ruiz-Ruano et al. 2016), so that fixation is improbable in

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

these conditions. In fact, although concerted evolution is the predominant pattern for satDNA evolution, non-concerted evolution has also been reported and explained through various factors such as life-history, population, location, organization, number of repeat-copies, or functional constraints (for review, see Garrido-Ramos 2015, 2017). However, the ultimate causes for concerted or non-concerted patterns are still unknown.

In this paper, we compare the full catalogue of satDNA families (i.e., the satellitome) between two grasshopper species belonging to the subfamily Oedipodinae, *Locusta migratoria* (Lmi) and *Oedaleus decorus* (Ode), which diverged 22.81 Ma (Song et al. 2015). We show the presence of about one-third of orthologous satDNA families whose sequence comparison pointed to mutation and drift as the main drivers of satDNA evolution. We also got estimates of nucleotide turnover rate at the level of consensus sequences (consensus turnover rate, CTR), using 20 orthologous pairs present in both species, and found that they were highly variable and depended on the history of satDNA amplifications. We also analyzed repeat landscapes and developed indices for satDNA homogenization and degeneration and an index for concerted evolution, which may be useful for future research. Also, we propose a general model for satDNA evolution and suggest that the evolution of these sequences constitute a good example of contingent evolution (see Blount et al. 2018).

Results

One-third of satDNA families showed sequence homology between species

The range of variation for repeat unit length (RUL) was 8-400 bp for the 60 satDNA families found in *L. migratoria* and 12-469 bp for the 58 families found in *O. decorus*. For subsequent analyses we included only those satDNA families showing more than 100 copies, which excluded the four least abundant satDNAs in *L. migratoria* (Table

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

S1). After comparing the consensus sequences of all satDNA families present in both species, we found that 21 families in *O. decorus* showed homology with 20 in *L. migratoria* (Table S2). We assume that these sets of satDNAs showing some sequence identity were already present in the most recent common ancestor of these two species (dated about 22.81 Ma) and thus belonged to the ancestor satDNA library. Therefore, these homologous sets constituted 12 orthologous superfamilies (OSFs) including 31 and 44 subfamilies in *O. decorus* and *L. migratoria*, respectively (Table S2). On the other hand, the non-shared satDNA families (37 in *O. decorus* and 36 in *L. migratoria*) could have arisen *de novo* after both lineages split, or else they had got lost in one of the species.

Between species comparison of basic satellitome features (Table 1) revealed that shared satDNAs did not show significant differences between species for RUL, A+T content, and abundance, but divergence was lower in *L. migratoria*. However, the non-shared satDNAs showed higher RUL and abundance in *O. decorus*. Within species comparisons between shared and non-shared satDNAs failed to show differences in *O. decorus*. In *L. migratoria*, however, the shared satDNA families showed higher RUL, A+T content and abundance, and lower divergence, than the non-shared ones (Table 1). Taken together, these results revealed the presence of many satDNA families showing short monomers among the non-shared ones in *L. migratoria* which also showed lower A+T content and abundance, but higher divergence than those shared with *O. decorus*.

Tandem structure and association with other repetitive elements

The quantification of homogeneous and heterogeneous read pairs allowed estimating the degree of tandem structure (TSI) for each satDNA family (Table S1). The annotation of the heterogeneous read pairs allowed identifying other genomic elements adjacent to

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

satDNA (Table S3). This revealed that LmiSat03-195 (TSI= 99.7%) was associated with LINEs in 57 out of the 100 heterogeneous read pairs observed. However, only 2% of the 1,356 heterogeneous read pairs showed association with LINEs for its orthologous OdeSat02-204 (TSI= 95.9%), suggesting that association with LINEs occurred only in *L. migratoria*. Likewise, OdeSat17-176 and LmiSat02-176 showed association with Helitron TEs in 93% and 76% of the 2,379 and 1,356 heterogeneous read pairs observed, respectively. Bearing in mind that the sequence of the LmiSat02-176 repeat unit shows homology with Helitron TEs (Ruiz-Ruano et al. 2016), the high frequency of association with Helitron observed for OdeSat17-176 and the low TSI (11.1%) suggest that most units detected for this satDNA were part of the TE itself and are not in tandem (i.e., 1-TSI= 88.9%). However, LmiSat02-176 showed high TSI (94.7%) and lower association with the TE (76%), suggesting that this satDNA arose from this TE, but it also constitutes an independent entity which has reached quite long arrays in *L. migratoria* (longer than 20 kb in the MinION reads). The FISH pattern of both satDNAs (see below) reinforced this conclusion, as OdeSat17-176 yielded no hybridization signals (Table 2), whereas LmiSat02-176 showed pericentromeric bands on six chromosome pairs (see Ruiz-Ruano et al. 2016 and Table S1).

A same orthologous satDNA may show different FISH patterns at intra- and interspecific levels

FISH analysis for 14 OdeSat families which showed homology with 20 LmiSat ones, revealed that six OdeSats displayed conspicuous bands on chromosomes (B-pattern from hereafter) whereas the eight remaining failed to show FISH signal (NS-pattern from hereafter), of which seven showed the B-pattern in *L. migratoria* (Table 2). This

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

revealed that a same OSF may show FISH signals in one species but not in a close relative.

To search for molecular differences between satDNAs showing the B- and NS-patterns, we analyzed MinION long reads in *L. migratoria* to score the maximum array length (MAL) for each LmisatDNA (Table 2). Even though coverage was very low (0.02x), we found that none of the seven NS families analyzed showed arrays higher than 2,500 bp, whereas almost half of those showing the B pattern did it (Gardner-Altman unpaired mean difference= 2930, 95.0%CI: 1540, 4790), and the three orders of magnitude of the difference indicated that satDNAs with the B-pattern have been submitted to more (and extensive) amplification events than those showing the NS-pattern. This difference justifies using the presence of FISH signals as an indication of the degree of satDNA amplification. The fact that 18 out of 20 orthologous satDNA families in *L. migratoria* showed the B-pattern, whereas only six out of the 14 orthologous families analyzed in *O. decorus* showed it, represent the first indication for a higher incidence of satDNA amplifications in *L. migratoria* (RxC contingency test, with 50,000 replicates: P= 0.00562, SE= 0.00077). This result was reinforced by the fact that the 14 OdeSat families included 24 subfamilies whereas the 20 LmiSat ones included 44 subfamilies (Table 2) (Wilcoxon matched-pairs test: z= 2.11, N=12, P= 0.035). As subfamilies represent different amplification events, the former results demonstrate that a same orthologous satDNA may show different amplification trajectories during their independent evolution in different species.

Careful examination of orthologous satDNAs revealed a unique case of no satDNA amplification in both species during the 22.8 Ma of separate evolution, as the LmiSat27-57 and OdeSat41-75 OSF showed the same NS-pattern. Consistently with their low degree of amplification, these two satDNAs showed very low values for TSI

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(9% in *O. decorus* and 32% in *L. migratoria*) and RPS (29% and 32%, respectively), indicating poor tandem structure and homogenization (see Tables 2 and S4). The remaining OSFs, however, showed amplification in at least one species. An interesting case was OSF7, where one of the five *L. migratoria* families showed the NS-pattern (LmiSat24-266) whereas the four remaining (LmiSat28-263, LmiSat43-231, LmiSat45-274 and LmiSat54-272) showed the B- pattern. Likewise, one of the two *O. decorus* families (OdeSat28-276) showed the B-pattern whereas the other (OdeSat58-265) showed the NS one. This shows that homologous satDNAs can display the NS or B patterns at intra- and interspecific levels. In fact, seven orthologous satDNA families with the NS-pattern in *O. decorus* showed the B-pattern in *L. migratoria* (Table 2).

One of the most dramatic differences was found for the orthologues OdeSat59-185 and LmiSat01-185, which were the scarcest and the most abundant satDNAs in *O. decorus* and *L. migratoria*, respectively, with the latter showing pericentromeric FISH bands on all chromosomes (Ruiz-Ruano et al. 2016) and OdeSat59-185 showing the NS-pattern (Table 2). Finally, even those satDNAs with FISH bands in both species showed remarkable differences regarding chromosome location (proximal, interstitial, or distal; see Table S1). Taken together, these results show that orthologous satDNAs can display disparate chromosome distribution in separate species due to their independent evolution. These differences can range from short arrays being undetectable by FISH, which may eventually serve as seeds for species-specific amplification (as suggested by Ruiz-Ruano et al. 2016), up to long arrays yielding conspicuous FISH bands.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SatDNA homogenization and degeneration

SatDNA evolution is debated between homogenization (through amplification) and degeneration (resulting from mutational decay). It would thus be desirable to find satDNA parameters being good indices for these two alternative states. To search for a homogenization index, we hypothesized that it should show a high negative correlation with intraspecific divergence. Spearman rank correlation analysis showed that, in both species, RPS (relative peak size, see methods and Fig. 1) showed a very high negative correlation with divergence (measured as K2P) (r_S= -0.9 in both species) (Table 3), which revealed RPS as a good homogenization index. On the contrary, a degeneration index should be negatively correlated with homogenization, and Spearman rank correlations revealed that DIVPEAK showed the highest negative correlation index with RPS in both species (Table 3). This means that the relative size of amplification peaks decreases as satDNA sequences accumulate divergence through mutational decay since the last satDNA amplification (see repeat landscapes in Figs. 2, S1 and Dataset 1).

To ascertain whether satDNA degeneration, measured by DIVPEAK, is associated with any of the satDNA parameters analyzed (RUL, A+T, no. subfam and TSI), we performed Spearman rank correlation analyses, which revealed that RUL was the only satDNA property showing significant correlation with DIVPEAK (Table 3) and it was negative and of similar magnitude as that between DIVPEAK and RPS. This suggests that RUL is an important determinant of satDNA degeneration, with shorter satDNAs degenerating faster. A possible explanation is that short monomers degenerate faster through mutational decay because every point mutation implies a higher proportion of degeneration for short than for long monomers, as if the Muller's ratchet would have fewer teeth for short than long repeat units and the same number of new

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mutations would imply a higher number of ratchet's turns for short repeating units than for long ones.

The analysis of the statistical properties of RPS and DIVPEAK indicated that, in both species, RPS fitted a normal distribution (ODE: χ^2= 4.45, df= 3, P= 0.215; LMI: χ^2= 4.78, df= 3, P= 0.189 whereas DIVPEAK fitted an exponential distribution (ODE: χ^2= 4.55, df= 2, P= 0.103; LMI: χ^2=4.93, df= 3, P= 0.177). Their scales ranged between 0 and 1 for RPS and between 0 and 27% (within the 0-40% scale of divergence measured here) for DIVPEAK.

We suggest that satDNA families follow evolutionary pathways that include recursive cycles of homogenization (through amplification by tandem duplication) and degeneration (through random mutation). After an amplification event, homogenization (measured by RPS) will increase, and degeneration (measured by DIVPEAK) will decrease. As time goes by, with no other amplification events, RPS will decrease and DIVPEAK will move towards higher values. An expected outcome of mutation accumulation is reducing the kurtosis of the repeat landscape (RL) distribution (i.e., curve flattening, Fig. 1 for examples). In fact, kurtosis was correlated negatively with DIVPEAK (Ode: N=58, r_S= -0.80, t= 9.89, P<0.000001; Lmi: N=56, r_S= -0.76, t= 8.58, P<0.000001) and positively with RPS (Ode: N=58, r_S= 0.80, t= 9.68, P<0.000001; Lmi: N=56, r_S= 0.83, t= 10.98, P<0.000001). Kurtosis is thus proportional to RPS, so that highly homogenized satDNAs show leptokurtic RLs whereas highly degenerated ones show platikurtic RLs. It is thus expected that kurtosis and RPS are high for recently amplified satDNAs and low for satDNAs which have not undergone amplification since long (see some examples in Figs. 2 and S1). Although these parameters do not constitute absolute measures of time, however, they can be useful as measures of "time

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

since the last satDNA amplification". As satDNA can undergo successive amplifications across evolutionary time, we can also consider RPS and kurtosis as rejuvenation indices.

To analyze whether conservation of the orthologous satDNA families in both species was associated with homogenization and degeneration indices, we compared them between the shared and non-shared satDNA families found in each species. In *O. decorus*, the effect size (unpaired mean difference) found between non-shared and shared satDNAs by means of Gardner-Altman estimation plots, revealed no mean differences for RPS (unpaired mean difference= -0.0682, 95.0%CI: -0.159, 0.0348), kurtosis (unpaired mean difference= 0.678, 95.0%CI: -1.62, 5.78) and DIVPEAK (unpaired mean difference= 1.13, 95.0%CI: -0.954, 5.61), indicating similar levels of homogenization/rejuvenation and degeneration in both groups. In *L. migratoria*, however, the three indices showed differences between shared and non-shared satDNA families, indicating higher homogenization/rejuvenation and lower degeneration for the shared ones (Fig. 3).

Amplification explains the concerted evolution of satDNA

O. decorus and *L. migratoria* shared their most recent common ancestor 22.81 Ma, on which basis we could perform estimations of interspecific rates of turnover in the consensus sequences (CTR). For this purpose, we compared the consensus DNA sequences of 20 pairs of orthologous satDNA, representing half of the 40 estimations that could be done at family level (see Table S2). The values obtained for CTR in the 20 orthologous pairs ranged from 0.013% (between LmiSat02-176 and OdeSat17-176) to 2.86% (between LmiSat03-195 and OdeSat02-204) nucleotidic changes in their consensus sequences per million year (mean= 1.11%, see Table 2), with two orders of magnitude between the extreme values.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To search for possible causes for such an extreme variation in the observed rates, we performed forward stepwise multiple regression of CTR (dependent) on four factors related to satDNA amplification: for each species, the number of subfamilies per satDNA family (subfam), the absolute number of copies included in the 5% divergence peak (peak-copies), RPS, and TSI. The results revealed that only three out of the eight factors entered a model that explained 85% of the total variance in CTR, with Ode_subfam explaining 56.4%, Ode_peak_copies explaining 25.7%, and TSI_Ode explaining only a nonsignificant 2.8% (Table 4). Variance inflation factors of this regression analysis ranged between 1.07 and 3.01 indicating the absence of multicollinearity. Likewise, the standardized residuals of this regression fitted a normal distribution (Shapiro-Wilks test: W= 0.97, P= 0.82). Finally, partial correlations were 0.85 for Ode_subfam, 0.76 for Ode_peak_copies, and 0.40 for TSI_Ode, whereas they were much lower for the five factors failing to enter in the model (from -0.25 to -0.02).

As we defined satDNA subfamilies by sharing 95% or higher sequence identity, i.e., up to 5% divergence, which was exactly the same figure used to define RPS and DIVPEAK on RLs, we consider that the number of subfamilies actually represents the number of independent amplification events being apparent within each family, as it also coincides with the number of different consensus sequences per family. As peak-copies represents the total number of repeat units in the amplification peak, we can infer that the rate of nucleotide change estimated from consensus sequences (CTR), which is positively correlated with the two former parameters, roughly represents the rate of nucleotide changes driven by satDNA amplification to be part of the consensus sequence. It was remarkable that only *O. decorus* variables entered in the stepwise multiple regression model, as it is the species showing the lowest number of subfamilies (31 versus 44 in the 12 OSFs, as a whole, and 24 and 44 in the 14 orthologous pairs

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

analyzed) and thus showed fewer amplification events, suggesting that CTR value is limited by the species showing fewer amplification events. We thus conclude that the same molecular mechanism, i.e., satDNA amplification, causes intraspecific homogenization and interspecific diversification, thus explaining the concerted evolution pattern of satDNA. Finally, we suggest that satDNA amplification (i.e., through unequal crossover leading to tandem duplication, i.e. a mutational mechanism) is the true force for Dover's *molecular drive*, but given its contingent nature, it should instead be named "*molecular drift*".

Most satDNA families showed concerted evolution in both species

Concerted evolution predicts that CEI>0, and this was met for 16 orthologous pairs, the four exceptions being the OdeSat17-LmiSat02 pair and three satDNA families in *O. decorus* (OdeSat41, OdeSat57, and OdeSat59) where CEI<0 thus showing signs of non-concerted evolution (Table 2). Remarkably, these four OdeSats failed to display FISH bands, suggesting that poor amplification might be related with non-concerted evolution. In both species, CEI was positively correlated with RUL (Ode: r_S= 0.70, N= 14, t= 3.4, P= 0.0051; Lmi: r_S= 0.56, N= 20, t= 2.83, P= 0.011) and RPS (Ode: r_S= 0.73, N= 14, t= 3.67, P= 0.0032; Lmi: r_S= 0.68, N= 20, t= 3.88, P= 0.0011) but not with A+T content (P>0.05 in both species). In addition, CEI was positively correlated with TSI in *O. decorus* (r_S= 0.78, N= 14, t= 4.26, P= 0.0011) but not in *L. migratoria* (r_S= 0.43, N= 20, t= 2.04, P= 0.056). Finally, in *O. decorus*, CEI was higher in the six satDNAs showing the FISH B-pattern than in the eight showing the NS-pattern (unpaired mean difference= 2.63; 95% CI: 0.883, 5.36).

These results indicate that satDNAs displaying longer monomers, higher levels of homogenization and the FISH B-pattern show higher indices of concerted evolution.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Exceptional non-concerted patterns were observed for satDNA families showing a low number of amplifications since all showed a single subfamily in both species.

The persistency of satDNA in these two species was not associated with functional constraints

Several sequence features have hitherto been associated with a variety of putative satDNA biological roles, the most relevant being centromere function. We searched for short internal repeats within each satDNA family's consensus sequences since these repeats have been associated with sequence function. We found no direct repeats within the sequence span of any satDNA sequence. On the contrary, it was common to find short inverted repeats in all satDNA families that might facilitate non-B DNA conformations such as stem-loops and cruciform structures, but they were found in both shared and non-shared satDNA families.

To ascertain whether Gibbs free energy (dG) of satDNA sequence depends on some satDNA properties, we performed forward stepwise regression, in each species, with dG as dependent variable and RUL, A+T, sharing status and degeneration status (DIVPEAK) as independent factors. In Ode, the regression model explained 67% of the variance in dG (59% by RUL, 5% by A+T, and 3% by DIVPEAK). The correlation was negative with RUL and positive with the two other factors. In *L. migratoria*, the result was highly similar, except that DIVPEAK did not enter in the model, but the dG variance explained was higher, reaching 83% (79% by RUL and 4% by A+T). As higher free energy values correspond to lower dG values, the former results indicate that free energy of satDNA sequence depends positively on RUL, as it determines the likelihood of autopairing, and, at lower extent, also depends on two other sequence properties influencing the number of hydrogen bonds in the double helix, as higher A+T

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

content implies more A-T pairs and fewer hydrogen bonds, thus lower free energy, whereas higher DIVPEAK indicates higher mutational decay that might difficult autopairing thus decreasing the number of hydrogen bonds. The fact that DIVPEAK of the shared satDNAs was higher in *O. decorus* than *L. migratoria* (paired mean difference= 2.6, 95.0%CI: 0.55, 6.8) is consistent with their higher degeneration in *O. decorus*.

We found that most of the shared satDNA families failed to show a propensity to acquire stable curvatures (Table S1), even though the curvature-propensity plots contained a peculiar maximum in some of them. However, the magnitude of these peaks (11 to 13 degrees/10.5 bp helical turn) was far from the values calculated for other highly curved motifs (Goodsell and Dickerson 1994; Gabrielian et al. 1996). Most intriguingly, these peaks were similar for satDNAs showing the NS or B FISH patterns or, in the latter case, whether they were located on pericentromeric regions or not. In total, only 11 (7 in *L. migratoria* and 4 in *O. decorus*) out of the 34 shared satDNA families showed curvature propensity, all showing RUL≥185 bp. They belonged to five different OSFs, three of which showed curvature propensity in both species, whereas the two remaining showed it in only one species, suggesting that this property does not depend only on RUL, which was highly similar in both species for these satDNA families.

We also analyzed curvature propensity for the non-shared satDNAs, and none of them showed it to a large degree. Notwithstanding, as observed for shared satDNAs, a few families (one in *L. migratoria* and five in *O. decorus*) showed a conspicuous peak of magnitudes between 11 to 14 degrees/10.5 bp helical turn. It has been suggested that DNA curvature may be involved in the recognition of DNA-binding protein components of the heterochromatin (Plohl et al. 2012). Our results show that curvature

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

propensity is not differentially frequent or relevant in the 34 shared satDNAs analyzed in both species, compared with the non-shared ones. Therefore, we believe that curvature propensity is not a relevant feature of satDNA or the cause for satDNA conservation in these two species.

Finally, we searched for the presence of short sequence motifs common to the shared satDNA families in both species. We isolated individual monomers from each satDNA family and calculated nucleotide diversity (π) per position (not shown). We did not find conserved motifs in these satDNAs, irrespectively of their FISH pattern or chromosomal location.

Taken together, these results show that, in these two species, there is no sequence conservation for pericentromeric satDNAs, which also lack significant sequence signatures other than A+T richness and repeat length. On the other hand, all putative functional signatures analyzed here were not more frequent in the shared satDNAs than in the non-shared ones. We interpret this as evidence that satDNA conservation is mostly a contingent event.

Incomplete sorting of the satDNA library

The satellitomes of relative species show sequence homology for a fraction of their satDNA families, which is the best support for the satDNA library hypothesis (Fry and Salser 1977). Joint analysis of RLs and MSTs revealed interesting properties of the satDNA library (Figs. 2 and S1): i) OdeSat02A and LmiSat03A were the two OSF02 subfamilies showing the highest amplification peaks in the RLs (Fig. 2a, plot on the left), and they also showed the highest CTR observed among all those analyzed here (2.86% per Ma). Remarkably, the MST plot for all subfamilies and families comprising OSF02 revealed complete sorting per species for this component of the library (Fig. 2a,

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

right). ii) On the other hand, OSF12 included two families in *L. migratoria* (LmiSat01 and LmiSat13) which were fully sorted in the MST (Fig. 2b, right), whereas the single *O. decorus* family (OdeSat59) was remarkably similar to LmiSat01A, with only two nucleotidic differences in their sequence, which is lower than those shown by the four other *L. migratoria* subfamilies with LmiSat01A. This illustrates an extreme case of incomplete library sorting (ILibS) and the second lowest CTR value (0.26% per Ma). Other OSFs showed intermediate situations. For instance, OSF04 showed CTR values between 1.16 and 1.60 and their MST revealed the existence of ILibS, with OdeSat32A being connected with three different LmiSats (37A, 26A and 51A), the latter being placed betwee OdeSat32A and OdeSat21A (see Fig. S1a). On the contrary, OSF5 (Fig. S1b) showed high CTR values (>2% per Ma) and complete library sorting, with the satDNAs properly separated between species. Finally, OSF07 showed CTRs between 0.56 and 1.43 and apparent ILibS, with high level of intermixing between the satDNAs of both species (Fig. S1c). Taken together, these observations suggest that CTR values are inversely associated with the level of ILibS. On this basis, we used the maximum CTR value (maxCTR= 2.86) as reference to estimate the degree of ILibS as one minus the quotient between CTR_i and maxCTR (see Table 2). This indicated that the satDNA library of *O. decorus* and *L. migratoria* shows, on average, 61% of incomplete sorting after 23 Ma. Finally, the fact that the four OdeSats showing the non-concerted pattern were those showing the highest ILibS figures (0.88-1), whereas ILibS values up to 0.84 corresponded with patterns of concerted evolution (see OSF8 in Table 2), suggested the possible existence of a threshold for ILibS (between 0.84 and 0.88) below which satDNA evolution is concerted.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Discussion

SatDNA evolution is mostly contingent

Comparative analysis of the satellitome in the grasshoppers *O. decorus* and *L. migratoria*, two species belonging to the Oedipodinae subfamily, which shared their most recent common ancestor about 23 Ma, gave us a chance to take a look into satDNA library evolution during this period. We assume that the 41 satDNA families (20 in *L. migratoria* and 21 in *O. decorus*) that showed sequence homology between species belong to 12 orthologue groups already present in the ancestor library, which have been conserved up today. However, the remaining 84 families (36 in *L. migratoria* and 37 in *O. decorus*) could represent either remnant satDNAs conserved in only one species or satDNAs arisen *de novo* during the separate evolution of these species. To distinguish between these two possibilities, it is necessary to analyze other oedipodine species. The occurrence of a species-specific profile of satDNAs resulting from differential amplifications and/or contractions from a pool of sequences shared by related genomes is a prediction of the library hypothesis of satDNA evolution with the subsequent replacement of one satDNA family for another in different species (Fry and Salser 1977). By analogy with incomplete lineage sorting (ILS) in phylogenetic studies, satDNA amplifications and/or contractions between close relative species may yield a pattern of incomplete library sorting (ILibS). We have detected here this phenomenon using consensus sequences, but the use of physical sequences would yield even higher rates of ILibS.

The library hypothesis predicts the residual retention of low-copy counterparts of the dominant satDNA of one species in the other (Fry and Salser 1977). For instance, OdeSat02A-204 and LmiSat03A-195 have been independently amplified in both species, reaching among the highest genomic abundances in both species, and showed the

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

highest CTR and extensive diversification, with four subfamilies in *O. decorus* and six in *L. migratoria* (see Fig. 2a). In addition, a joint MST for OSF02 (to which both satDNA families belong) revealed the absence of ILibS as all satDNA families and subfamilies appeared well separated between species in the MST (see Fig. 2a). Conversely, the consensus sequences of LmiSat01A-185 and OdeSat59-185 only differed in two positions, thus showing higher interspecific similarity than that found, at intraspecifical level, between the five *L. migratoria* subfamilies (see Fig. 2b), thus constituting an extreme example of ILibS. The high similarity in the consensus sequences of OdeSat59A and LmiSat01A cannot be explained by functional conservation because only the latter shows FISH bands on centromeric regions of all chromosomes thus probably playing a centromeric function in *L. migratoria*, whereas OdeSat59A is the most scarce satDNA found in *O. decorus* thus being only a relic. Likewise, while OdeSat01-287 is the most abundant satDNA in *O. decorus*, its orthologous (LmiSat09-181) is a relict in *L. migratoria*. We thus believe that the observed sequence similarity between OdeSat59A and LmiSat01A might be due to chance convergence, as the likelihood of nucleotide coincidence in each position of the consensus sequence is a function of the relative frequency of the four possible nucleotides in each species, thus being a probabilistic issue.

Our estimates of ILibS from CTR values indicated that the satDNA libraries of *O. decorus* and *L. migratoria* still show 61% of incomplete sorting after 23 Ma of independent evolution, i.e. about 39% of complete sorting (1.7% per Ma). This extreme cohesiveness of the satDNA library is due to the highly paralogous nature of these genomic elements, with thousand copies evolving at once, independently in both species, through point mutation, amplification (tandem duplication) and drift (see below). This 39% expresses only part of library divergence, as the maximum divergence would be

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

523 reached when all homology signals between satDNAs in both species would have been
524 erased, as in the case of the non-shared ones, whereas the satDNAs belonging to OSF02
525 are still recognized as homologous between species even with 100% library sorting.
526 Anyway, the ILibS parameter of a given OSF (or orthologous pair of satDNAs)
527 inversely indicates its possible utility for phylogenetic analysis.

528 Another prediction of the library hypothesis is that the appearance of satDNA
529 families would usually represent amplification of one of the satellites already present at
530 a low level in the library, rather than actual *de novo* appearance. It is not easy to know if
531 any of the non-shared satDNA families actually arose *de novo*. However, in *L.*
532 *migratoria*, the lower RUL of non-shared satDNAs suggests that the satellitome of this
533 species might harbor some *de novo* arisen short satellites, in consistency with an
534 evolutionary trend towards increasing monomer length and complexity, suggested by
535 theoretical (Stephan and Cho 1994) and experimental (Garrido-Ramos et al. 1995; de la
536 Herrán et al. 2001a; Navajas-Pérez et al. 2005; Ruiz-Ruano et al. 2018a) work.

537 Our estimates of CTR by the comparison of 20 orthologous pairs of satDNA
538 families indicated that it was 1.11% per Ma, which implies that two satellites can
539 diverge by more than 50% in about 50 Ma. This explains why *L. migratoria* and *O.*
540 *decorus*, belonging to the Acrididae family do not share a single satDNA family with
541 *Eumigus monticola* (Ruiz-Ruano et al. 2017), a grasshopper belonging to the
542 Pamphagidae family, as these two orthopteran families shared their most recent
543 common ancestor about 100 Ma (Song et al. 2015). Along with the stochastic nature of
544 satDNA loss or gain during evolution, sequence changes at the mentioned rate will
545 make unrecognizable a satDNA family after 100 Ma of separate evolution within the
546 genomes of different species, which contrasts with the case of some other satDNAs
547 preserved for more than 60 Ma (Garrido-Ramos et al. 1999; de la Herrán et al. 2001b;

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Mravinac et al. 2002; Cafasso and Chinali 2014) or even more than 100 Ma (de la Herrán et al. 2001a; Robles et al. 2004).

Our results suggest that the same OSF may be involved in the centromeric function in a given species but not in a close relative species. According to Melters et al. (2013), the most abundant satDNAs in a genome are most likely involved in the centromeric function. Another feature suggesting this fact is satDNA location on pericentromeric regions of all chromosomes. Therefore, LmiSat01-185, OdeSat01-287 and/or OdeSat02-204 are the best candidate families in these species since all meet the two conditions. However, all three satDNAs showed orthologous families in the other species displaying much more limited chromosome distribution, suggesting that one or both species have replaced the centromeric satDNA during the last 22.8 Ma. No significant track of signatures such as conserved motifs or sequence mediated specific stereo-spatial features were found for these or any other pericentromeric satDNAs found in these species. We thus believe that, in the absence of other evidence, contingent facts such as the opportunity to be in the right place when amplified might be responsible for centromeric satDNA turnover. Zhang et al. (2014) also revealed rapid divergence for centromeric sequences among closely related *Solanum* species and suggested that centromeric satellite repeats underwent boom-bust cycles before a favorable repeat became predominant in a species. Indeed, there are species such as chicken (Shang et al. 2010), common bean (Iwata et al. 2013), or pea (Neumann et al. 2012) that contain different satDNAs in different centromeres.

Whether a given satDNA is conserved for long due to functional reasons is an open question. Fry and Salser (1977) suggested that an essential step in the evolution of a specific satDNA family may be acquiring a biological function. However, persistence over time of a satDNA might also be explained in terms that do not depend on natural

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

selection (Stephan 1986, 1987, 1989; Walsh 1987; Harding et al. 1992). Our results were consistent with this latter view. No conserved functional motifs were found within the monomers of every grasshopper satDNA analyzed as has been found in other satDNAs such as human centromeric satDNA (Masumoto et al. 1989, 2004; Muro et al. 1992; Haaf et al. 1995). On the other hand, short dyad symmetries within satDNA repeats might be associated with thermodynamically stable secondary structures and yield non-B-form conformations, such as stem-loops or cruciforms. It has been claimed that these short dyad symmetries may play an important role in satDNA repeats as targets for protein binding and thus in satDNA function (Koch 2000; Hall et al. 2003; Luchetti et al. 2003; Plohl et al. 2012; Pezer et al. 2012; Garrido-Ramos 2015, 2017). Kasinathan and Henikoff (2018) have proposed that that cruciform structures formed by dyad symmetries may specify centromeres and that these non-B form DNA configurations in centromeric repeats may facilitate centromere assembly (Kasinathan and Henikoff 2018; Talbert and Henikoff 2018). In the two grasshopper species analyzed here, short inverted repeats that might facilitate dyad symmetries and non-B DNA conformations were frequent in both shared and non-shared satDNAs, independently of their organization and chromosomal location. We believe that this property is a simple outcome of stochastic processes of satDNA evolutionary dynamics. Its ubiquity suggests that almost any satDNA can be recruited for functions being dependent on the formation of non-B DNA conformations (see Kasinathan and Henikoff 2018).

SatDNA evolution is a topic of high interest for the scientific community, but with poor agreement about general pathways and mechanisms. Molecular drive was a turnover mechanism suggested by Dover (1982a,b) as a directional force for repeat fixation, in general, which has been the prevalent hypothesis for satDNA evolution due

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

598 to its apparent explicative power as a mechanism for sequence change, turnover, and
599 concerted evolution. Nonetheless, the presence of satDNA arrays on multiple genomic
600 sites makes it impossible, in practice, the fixation of a given satDNA repeat. The
601 positive association of the number and extent of satDNA amplifications with the
602 nucleotide substitution rate observed from consensus sequences suggests that molecular
603 drive is actually a mutational force (tandem duplication by means of crossing-over) able
604 to change copy numbers among the different sequence variants pre-existing for a given
605 repeat family, most frequently leading to incomplete turnovers, and it operates mainly
606 through satDNA amplification. A good way to visualize the role of amplification in
607 satDNA evolution is through repeat landscapes for families consisting of several
608 subfamilies remaining at low frequency for high divergence values that, at lower figures,
609 show amplification peaks for one or more subfamilies (see Figs. 2 and S1).

610 The high or low degree of homogenization for a given satDNA is inversely
611 proportional to the time since the last amplification. It thus depends on i) the neutral
612 mutation rate introducing new sequence variants (increasing intra-specific divergence)
613 and ii) the rate of satDNA amplification, implying partial turnovers that promote
614 sequence variants that become new subfamilies. As satDNA amplification for
615 orthologous satDNA families is independent in relative species, it behaves as an inter-
616 specific drifting mechanism. This dual role of satDNA amplification as the major
617 homogenizing force at the intraspecific level and as the principal driver for interspecific
618 sequence divergence, forced by reproductive barriers, inevitably leads to the concerted
619 evolution pattern. In fact, 16 pairs of orthologous satDNAs met this pattern, with only
620 four showing a non-concerted one. Remarkably, these exceptions coincided with the
621 absence of major amplifications in *O. decorus* satDNAs that remain at low abundance.
622 This kind of variation can persist for long in the absence of (homogenizing)

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

amplification events (Navajas-Pérez et al. 2009). Therefore, concerted evolution should be a reasonable consequence of the stochastic nature of satDNA evolution, while exceptional non-concerted patterns can result from differential amplifications among species. Other exceptions can result from satDNA homology with TEs, as was the case for LmiSat02-176, whose homology with Helitron might have biased the calculation of intraspecific divergence. Other explanations have been raised as possible causes for non-concerted evolution patterns, such as the effect of location, organization, and repeat-copy number (Navajas-Pérez et al. 2005, 2006, 2009), population and evolutionary factors (de la Herrán et al. 2001a; Robles et al. 2004; Suárez-Santiago et al. 2007; Quesada del Bosque et al. 2013, 2014), biological factors (Luchetti et al. 2003, 2006; Lorite et al. 2017), or functional constraints (Mravinac et al. 2005).

We have shown here that concerted evolution is a pattern emerging from satDNA amplification due to the resulting homogenization at intraspecific level and diversification at interspecific level. To visualize this relationship, think about two species recently emerged from a common ancestor. Their satDNA libraries are almost identical at interspecific level but both retain the ancestral polymorphism at intraspecific level. This situation would imply, for each OSF, ILibS values next to 1 and CEI<0 since divergence would be higher at intra- than inter-specific level. As time goes by and mutation and drift operate, ILibS will decrease and CEI will increase as new mutations occur independently in both species. In absence of satDNA amplification, mutation and drift would lead satDNA towards concerted evolution by increasing interspecific divergence, although this process would be slow. However, the pathway to concerted evolution would be paved away by satDNA amplification as the resulting homogenization would reach CEI>0 values (by sharply decreasing intraspecific divergence) when ILibS would decrease below a threshold which, in the case of *O.*

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

decorus and *L. migratoria*, lies between 0.84 and 0.88. The fact that this threshold is so close to 1 reinforces the idea that concerted evolution is an unavoidable property fastly emerging from satDNA amplification. In fact, the four satDNA families which in *O. decorus* showed signs of non-concerted evolution showed low levels of homogenization (RPS between 0.29 and 0.40) and high values of ILibS (0.88-1), presumably due to the low level of amplification of these four satDNAs in this species. Taken together, our results indicate that concerted evolution is a state of interspecific diversification of the satDNA library, reached below a given ILibS threshold, which is fastly promoted by satDNA amplification.

A model for satDNA evolution

Considering all findings derived from the quantitative analysis of 114 satDNAs in *O. decorus* and *L. migratoria*, we suggest the following model for satDNA evolution (Fig. 4). Intragenomic changes are mainly stochastic, implying that satDNA families mainly evolve under the domain of mutation and drift. SatDNA arises from any tandem duplication yielding at least two monomers. Subsequent unequal crossover is the main source for longer arrays with the consequent increase in tandem structure. This tandem duplication is one of the two classes of mutation operating on satDNA. The other is point mutation increasing divergence among the different monomers composing the whole set of satDNA sequences belonging to a given family. When tandem duplication occurs massively during a short time, it constitutes an **amplification** event that decreases intra-specific divergence (i.e., increases homogenization as measured by RPS) by adding a high number of repeats showing identical sequence. Next, intra-specific divergence will grow across years by the incidence of point mutations, inevitably leading to the **degeneration** of the satDNA sequence unless new amplifications occur.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

This is characterized by a temporal decrease of RPS and kurtosis and an increase of DIVPEAK as family sequences became more and more divergent. From time to time, some monomers will lose their identity as members of a given satDNA family (reaching identities lower than 80%) or even as members of the same superfamily (with no recognizable homology). This process may shorten long arrays into pieces, thus decreasing TSI and, finally, the satDNA may fade away across time.

In a sense, every new amplification event **rejuvenates** the satDNA family by promoting a given subfamily to the highest abundance and homogenization, after which it begins its degeneration process until new amplifications rejuvenate it again, or else fades out through accumulation of point mutations. In summary, we suggest that satDNA undergoes recursive cycles of amplification-degeneration-rejuvenation that may keep them in the genome for a long time. During this time, they can integrate into longer repeat units or higher-order structures (Willard and Waye 1987; Warburton and Willard 1990), or else disappear through sequence degeneration. The fact that short satDNAs degenerate faster than the longer ones (see above) suggests that their cycle is usually shorter than that of long satDNAs, partly explaining why many short satDNAs show high K2P divergence and platykurtic distribution. For instance, LmiSat10-9 is made of monomers of only 9 bp and is not found in Ode. Even if it would have been present in the common ancestor, it is doubtful that it would have remained for 22.8 Ma in both species without losing identity in at least one of them. In fact, there seems to be a minimum monomer length for homology conservation in these two species, which was 57 bp (LmiSat27-57 and OdeSat41-75). Alternatively, a satDNA formed by repeats of only 9 bp could have arisen *de novo*, by chance, in the gigantic genome of *L. migratoria* (Ruiz-Ruano et al. 2016).

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In addition to all former intragenomic events, satDNA frequently undergoes spread among chromosomes. Transposition and replication of extrachromosomal circles of tandem repeats, by the rolling-circle mechanism, followed by reinsertion of replicated arrays, have been postulated as the main mechanisms for the amplification and spread of satDNA families and is supported by indirect (Fellicielo et al. 2005, 2006) or direct (Cohen et al. 2005, 2010) evidence.

At intergenomic (population) level, the only conceivable way to spread an amplification event (occurred in a single individual) is through differential reproduction, as we believe that the molecular drive mechanism suggested by Dover (1982a,b) as a non-selective fixing force even at the population level, is circumscribed at the intragenomic level. Differential reproduction can occur at random, i.e., by genetic drift, or non-random, i.e., through selection. The latter may be negative, setting up an upper limit to the amount of satDNA tolerable by a genome. Purifying selection, mutation and drift are the drivers in the mutational-hazard (MH) hypothesis (Lynch 2011; Lynch et al. 2011), which suggests that the efficacy of purifying selection is impaired by genetic drift in small populations. This is especially applicable to satDNA, where CTR is highly variable among families (intragenomically). The fact that all satDNA families within a genome have been submitted to the same demographic changes at population level (excepting the differences due to sex linkage) means that purifying selection appears to set few limits to the variation in nucleotide substitution rate among satDNA families. Interestingly, 18 out of 20 shared satDNA families in *L. migratoria* showed amplification events giving rise to FISH bands, whereas only six out of their 14 orthologous families in *O. decorus* did it. This reveals that many of these OSFs have shown highly different evolutionary paths in both species. Based on the MH hypothesis, we may speculate that the extreme demographic changes associated with locust

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

outbreaks in *L. migratoria* might have helped to spread individual satDNA sequences at the population level during the extreme bottlenecks that characterize the solitary phase and subsequent population expansions during the gregarious one. This issue needs further research, including quantitative population analyses of every satDNA family in this species.

In addition, selection can operate positively through non-phenotypic (i.e., meiotic drive) or phenotypic (functional recruitment) effects, as is the case for centromeric and telomeric repeats. The latter is the extreme example of functional recruitment since the repeat is actively homogenized by an RNA-protein complex (telomerase) coded by the genome. Centromeric satDNA in primates resembles this kind of recruitment as another gene (CENPB) is involved in the organization of centromeric satDNA (Masumoto et al. 1989, 2004; Muro et al. 1992; Haaf et al. 1995).

Our model is an extension of the models devised in the '70s and '80s (Kimura and Ohta 1979; Ohta 1981, 1983; Ohta and Kimura 1981; Stephan 1986, 1987, 1989; Charlesworth et al. 1986), with some more emphasis on the intragenomic level, and under the light of the MH hypothesis (Lynch 2011; Lynch et al. 2011). Briefly, amplification is the homogenizing force, point mutation causes sequence degeneration, and new amplifications rejuvenate satDNA. We believe that our model brings about some essential term clarifications. For instance, Escudeiro et al. (2019) recently suggested a model of satDNA evolution in bovids consisting of three stages, namely amplification, degeneration (deduced from high satDNA similarity between some species and low between others) and homogenization (high sequence identity among all species). These authors thus claimed for degeneration and homogenization as if they were inter-specific processes. However, in our model, both processes are intragenomic (i.e., intra-specific) resulting from satDNA amplification and point mutation,

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

respectively, whereas inter-specific homogenization or degeneration is highly unlikely under contingent evolution. In fact, homogenization to an identical sequence in several species could only be achieved by functional (selective) recruit, as that occurred for the telomeric DNA repeat.

Finally, the paralogous nature of the satDNA library implies that its diversification between species is much slower than that of single-copy DNA, with high levels of incomplete library sorting which may be a problem for the use of satDNA for phylogenetical purposes beyond satDNA evolution itself. However, the pathway followed by an ancestor satDNA library after speciation can be monitored by satellitome comparison, as shown here for *O. decorus* and *L. migratoria*. A new body of research is taking form recently about contingency and determinism in evolution (Blount et al. 2018), trying to answer Gould's question on whether evolutionary trajectories are repeatable (Gould 1989). In this respect, satellitome evolution is a natural "parallel replay experiment" able to show many properties of contingent evolution, as the initially identical libraries in the ancestor undergo independent evolution after speciation reaching a high diversity of outcomes among different OSFs. Within species, the environment (at both intragenomic and population levels) is the same for all satDNA families (except for genomic location and organization), but the pathway followed by each of them is highly variable: some families show consensus sequences being highly similar to those in the other species, thus showing high ILibS, whereas others are completely sorted between species, and still others are unrecognizable between species because they have arisen *de novo* in one species or else they have undergone so many sequence changes that have lost homology between species. In analogy with Blount et al. (2018) claiming at ecological level, the evolutionary trajectory followed by each OSF in the satellitomes of two separate species

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

is mainly influenced by stochastic processes (i.e. mutation and drift), most likely reaching different outcomes even when both species satellitomes started from the same state in the ancestor and the different OSFs evolved under almost identical conditions at intragenomic level. Therefore, the satellitome is a good example of contingent evolution supporting that "disparate outcomes become more likely as the footprint of history grows deeper" (Blount et al. 2018). A rough estimate of the minimal degree of contingent evolution in the *O. decorus* and *L. migratoria* satellitomes can be obtained from the 20 orthologous satDNA pairs used here to estimate CTR. As Table 2 shows, only two of them showed identity higher than 95%: OdeSat17-176/LmiSat02-176 showing a single nucleotide difference in their consensus sequences, and OdeSat59-185/LmiSat01A-185 showing two differences. The first pair showed homology with Helitron TEs which could have biased identity calculations, and the second one appears to have little to do with functional conservation (as explained above). Even assuming that these two cases are adaptive convergences (which is unlikely), we can estimate that satDNA evolution in these species was at least 90% contingent.

Methods

Materials and sequencing

We collected 21 males of the grasshopper *Oedaleus decorus* in Cortijo Shambala (Sierra Nevada, Granada, Spain; 36.96111 N, 3.33583 W) on 6 July 2015. They were anaesthetized with ethyl-acetate vapours prior to dissection, and testes were fixed in 3:1 ethanol-acetic acid and stored at 4ºC for subsequent fluorescent in situ hybridization (FISH) analysis. Body remains were immersed in liquid nitrogen and stored at -80 ºC for molecular analysis and DNA sequencing. We then extracted genomic DNA from a hind leg from one male, using the GenElute Mammalian Genomic DNA Miniprep kit

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(Sigma). Next we sent the purified DNA to Macrogen Inc. (South Korea) who built a genomic library with ~180 bp insert size, using the Illumina Truseq nano DNA kit, and sequenced it in an Illumina HiSeq2000 platform (2x101 nt) yielding about 9 Gb of reads. We deposited this library in the Sequence Read Archive (SRA) under accession number SRR9649806.

For the *Locusta migratoria* satellitome, we used the results generated in Ruiz-Ruano et al. (2016), including some new analyses of the same Illumina libraries obtained from a Spanish individual lacking B chromosomes (SRA library SRR2911427), satDNA FISH location, and their consensus sequences (GenBank accession numbers KU056702–KU056808). During these new analyses, we detected a previous mistake in the assembly of the LmiSat01A-193 subfamily, consisting of a false tandem duplication of 8 nt in the consensus monomer. We amended this mistake and renamed the (new) sequence as LmiSat01A-185 (GenBank accession number KU056702.2). We thus performed a new analysis of abundance and divergence for the whole satellitome, considering this modification that implied only slight changes.

In addition, we generated an Oxford Nanopore library for *L. migratoria* using the MinION system with a flow cell version R9. We constructed the library using 5 µg of DNA without fragmentation step applying the the Nanopore Genomic Kit version SQK-LSK108 and the CleanNGS magnetic beads for washes. After applying the localbase-calling program from Nanopore, we got 63,346 reads summing up 130 Mb (~0.02x of coverage).

Bioinformatic and sequence analyses

We characterized the *O. decorus* satellitome applying the satMiner protocol (Ruiz-Ruano et al. 2016). Briefly, this protocol begins with a run of RepeatExplorer (Novák et

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

al. 2013) and the elimination of homologous reads with Deconseq (Schmieder et al. 2011) to perform a new round of RepeatExplorer with the remaining reads. We started with 100,000 read pairs and performed five additional rounds, subsequently duplicating the number of read pairs. Then we identified clusters in each RepeatExplorer round showing spherical or ring-shaped graphs, which are typical for satDNA. We checked the structure of their contigs with a dot-plot using Geneious v4.8.5 (Drummond et al. 2010) to test if they were tandemly repeated, and only those that met this condition were considered as satDNA. Every satDNA family was named with three letters alluding to species name (*L. migratoria* or *O. decorus*) followed by "Sat", a catalogue number (in decreasing order of abundance) and monomer length, following our previous suggestion in Ruiz-Ruano et al. (2016). For instance, the most abundant satDNA families in the two species analyzed here were LmiSat01-185 and OdeSat01-287. The different subfamilies within a same family were alphabetically named with capital letters in order of decreasing abundance.

Considering their level of sequence identity, we classified every collection of homologous sequences into subfamilies (identity>95%), families (>80%), and superfamilies (>40%). Next, we randomly selected 5 million read pairs with SeqTK (https://github.com/lh3/seqtk) and aligned them against the reference sequences with RepeatMasker v4.0.5 (Smit et al. 2013). With these results, we estimated total abundance and average divergence and generated a repeat landscape. Finally, we numbered the satellite families in descending order of abundance. We deposited sequences for satellite DNAs characterized in *O. decorus* in GenBank with accession numbers MT009035 - MT009125.

We then searched for homology between *L. migratoria* and *O. decorus* satellitomes with the rm_homolgy script (Ruiz-Ruano et al. 2016) that makes all-to-all

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

alignments with RepeatMasker (Smit et al. 2013). We aligned homologous satellites with Muscle v3.6 (Edgar 2004) implemented in Geneious v4.8.5 (Drummond et al. 2010) and reviewed them manually. Then we generated minimum spanning trees (MST) with Arlequin v3.5 (Excoffier and Lischer 2010) and visualized them with HapStar v0.7 (Teacher and Griffiths 2011). We used the same alignments to estimate the divergence between satDNA families of *L. migratoria* and *O. decorus*. To estimate a consensus turnover rate (CTR) of satDNA sequences, we performed alignments of consensus sequences using ClustalX (Thompson et al. 1997). Sequence divergence between species was calculated according to the Kimura two-parameter model (K2P; Kimura 1980), using MEGA6 (Tamura et al. 2013). When orthologous satDNA families were composed of several subfamilies, all consensus sequences from each subfamily were aligned and the average of all pairwise distances between the two species was computed. Finally, CTR was calculated using the CTR= K/2T equation, where T= divergence time between species and K= K2P divergence (Kimura 1980). Turnover rates were estimated considering that the *Oedaleus* and *Locusta* genera split 22.81 Ma (Song et al. 2015).

To get some insights on array length, we analyzed our MinION library obtained from *L. migratoria* gDNA (see above). For this purpose, we performed an alignment of these reads against the consensus sequences of the *L. migratoria* satellitome using RepeatMasker (Smit et al. 2013). However, due to the lack of resolution at subfamily level due to the high level of sequencing errors in these long reads, we only performed this analysis only for the most abundant subfamily in each family, i.e, that noted with the letter “A”. We then analyzed the length of all arrays found for each family to recorded the maximum array length (MAL) for subsequent analysis. For this purpose, we only considered arrays showing length higher than 1.5 repeat units, i.e. at least dimers, and the observed figures for MAL in the 56 satDNA families analyzed in *L.*

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

migratoria ranged between 62 and 20,180 repeat units. In addition, we considered 3 nt as the maximum inter-array distance to collapse two consecutive TR arrays into a same array, in order to partly counteract the splitting effect of short insertions or deletions due to replication slippage. These calculations were implemented in a custom script (https://github.com/mmarpe/satION/blob/master/dis_bed_max.py).

Analysis of tandem structure

We developed a method to estimate the degree of tandem structure in satDNA using a pipeline that we made publicly available throughout repository (https://github.com/fjruizruano/SatIntExt). This method is based on scoring the number of Illumina read pairs containing repeat units for a given satDNA family in the two reads (onwards named "homogeneous read pairs") and the number of read pairs containing such a repeat in only one member of the read pair (onwards named "heterogeneous read pairs"). The proportion of homogeneous read pairs indicates the degree at which a satDNA family is tandemly structured (tandem structure index = TSI). This index underestimates the true value by the equivalent to the half of the number of arrays (since each array has two external units). However, as the number of repeat units is much higher than the number of arrays, we consider that this underestimation may be low at the genomic level. To validate TSI, we analyzed Oxford Nanopore MinION long reads in *L. migratoria*, by annotating all satDNA variants found in them and scoring the number of repeat units constituting the longest array found for each satDNA family. Despite low coverage of the MinION reads, these longest arrays showed significant positive correlation with TSI (Spearman rank correlation: r_S= 0.42, N= 55, t= 3.36, P= 0.001), indicating that TSI is a valid estimator for the degree of tandem structure of satDNA. In addition, we tried to annotate the external read of every heterogeneous read

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

pair with the database of repetitive elements of *L. migratoria* generated in Ruiz-Ruano et al. (2018b) with RepeatMasker. Thus, we found homology of the elements adjacent to the satDNA arrays with satDNAs, transposable elements, rDNAs, snDNAs, tRNAs, histones, mitochondrial DNA and unknown elements in some read pairs, and counted the number of occurrences. This analysis is also integrated in the above-mentioned pipeline.

Homogenization and degeneration indices

SatDNA homogenization, i.e., the degree of intraspecific similarity between its tandemly structured monomers, is conceptually inverse to average sequence divergence. Therefore, a homogenization index should be negatively correlated with the K2P divergence. Trying to get such an index, we built repeat landscapes for each satDNA subfamily (90 in *O. decorus* and 103 in *L. migratoria*) and searched for divergence peaks, i.e., those divergence values showing the highest abundance in the repeat landscape (DIVPEAK) (Fig. 1). Then, we summed up the abundances of all satDNA sequences at ±2% divergence from the DIVPEAK class to calculate abundance in the 5% peak or PEAK-SIZE (Fig. 1). The logic was to get a collection of sequences diverging 5% or less to the consensus sequence, thus coinciding with our criterion to define subfamilies, as they probably derived from the same amplification event (see Ruiz-Ruano et al. 2016 for details). Finally, we calculated relative peak size (RPS) as the quotient between PEAK-SIZE and total abundance (see Fig. 1), which measures the proportion of repeat units being part of the last amplification event. To calculate RPS at the family level in those families showing two or more subfamilies, we followed the same procedure including all subfamily satDNA sequences, so that each subfamily weighted in proportion to its abundance. RPS serves as an index of homogenization

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

because it is expected to increase with satDNA amplification, as the new units derived from tandem duplication will initially show identical sequences, thus increasing global identity. DIVPEAK serves as an index of degeneration because it will increase by mutation accumulation and is thus proportional to the time passed since the last amplification. Specifically, DIVPEAK is the value of divergence (from 0% onwards) at which a given satDNA shows its maximum abundance, and increases when mutational decay move its abundance peak away from complete homogenization (divergence=0) where it arrived after its last major amplification event. The values for average divergence, total abundance, maximum abundance, maximum divergence, RPS and DIVPEAK for every satDNA family were estimated from with a custom script using the divsum files from RepeatMasker (https://github.com/fjruizruano/SatIntExt/blob/main/divsum_stats.py).

Concerted evolution index and incomplete library sorting

We calculated the divergence at intra- ($K2P_{intra}$) and inter-specific ($K2P_{inter}$) levels for the 20 pairs of orthologous satDNA families, and calculated an index of concerted evolution (CEI) as log2 the $K2P_{inter}/K2P_{intra}$ quotient.

The comparative analysis of RLs and MSTs revealed that the observed differences between OSFs in CTR were due to the state of library sorting between species. On this basis, we observed that the OSF showing the highest CTR was that showing a best separation between species for all families and subfamilies of satDNA. We then gave 1 to the sorting state of this OSF and then divided all CTR values by this maxCTR to obtain an index of the relative sorting for each OSF. One minus the obtained value thus indicated the degree of incomplete library sorting (ILibS) for each OSF.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Analysis of conserved motifs and curvature

We analyzed the consensus sequences of shared and non-shared satDNAs between the two species looking for functional signatures. We used the ETANDEM, EINVERTED, and PALINDROME programs from the EMBOSS suite of bioinformatics tools (Rice et al. 2000) for the detection of internal repeats (direct or inverted) and palindromes. Short internal direct repeats indicate the presence of functional motifs within the satDNA repeats. Dyad symmetries, many of them associated with thermodynamically stable secondary structures, are predicted to adopt non-B DNA conformations, such as stem-loops or cruciforms, which might have a role as targets for protein binding. Thus, as an additional test on the propensity to form non-B DNA conformations, we checked all satDNA families using the Mfold web server (http://www.unafold.org/mfold/applications/rna-folding-form-v2.php) for nucleic acid folding prediction (Zuker 2003), estimating Gibbs free energy (dG) of the predicted secondary structures (SantaLucia, Jr 1998). We also checked the consensus sequences of both types of satDNAs for sequence-dependent bendability/curvature propensity of repeats. We produced the bendability/curvature propensity plots with the bend.it server at http://pongor.itk.ppke.hu/dna/bend_it.html#/bendit_intro (Vlahovicek et al. 2003), using the DNase I based bendability parameters of Brukner et al. (1995) and the consensus bendability scale (Gabrielian and Pongor 1996). Finally, we used the sliding windows option of the DnaSP v.5.10 program (Librado and Rozas 2009) for the analysis of nucleotide diversity (π) per position for every shared satDNA in order to detect DNA conserved motifs. For this, we use multiple alignments of several dozens of monomer repeats selected per each satDNA.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Chromosomal location of the O. decorus satDNAs

To compare the chromosomal location of orthologous satDNA families in these species, we performed fluorescent in situ hybridization (FISH) for 14 satDNA families in *O. decorus* which showed sequence homology with 20 families in *L. migratoria*. For this purpose, we designed divergent primers for these 14 satDNA families in *O. decorus* using Primer3 (Untergasser et al. 2012) with a Tm ~60 ºC, to generate FISH probes as described in Cabrero et al. (2003) and Ruiz-Ruano et al. (2016).

Statistical analysis

To investigate distribution fitting of RPS and DIVPEAK, we used the chi-square test, and the normality of other variable distributions was tested by the Shapiro-Wilks test, and, when this condition was not met, we used the non-parametric Spearman rank correlation test. In the case of turnover rate, we performed forward stepwise multiple regression to analyze its dependence on other variables. In this case, we calculated variance inflation factors (VIFs) to test for multicolinearity, and the fit of standardized residuals of this regression to a normal distribution was tested by means of the Shapiro-Wilks test. All these analyses were performed using the Statistica software (Statsoft Inc.). Two-group comparisons were performed by the Gardner-Altman estimation plot method devised by Ho et al. (2019) following the design in Gardner and Altman (1986), as implemented in https://www.estimationstats.com. This analysis calculates the effect size by the mean difference between groups, for independent samples, or else by the paired mean difference in case of paired samples. The effect size is then evaluated by the 95% confidence interval (95% CI) and whether it includes or not the zero value. Contingency tests were performed by the RXC program, which employs the Metropolis

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

algorithm to obtain an unbiased estimate of the exact p-value (Rousset and Raymond 1995). In all cases 20 batches of 2,500 replicates were performed.

Abbreviations

B-pattern: Banded pattern (pattern in FISH analyses)

CEI: Concerted Evolution Index

CI: Confidence Interval

CTR: Consensus Turnover Rate

dG: Gibbs free energy

DIVPEAK: Divergence Peak

FISH: Fluorescence *In Situ* Hybridization

ILibS: Incomplete Library Sorting

K2P: Kimura Two-Parameter (substitution model)

Lmi: *Locusta migratoria*

NS-pattern: No signal pattern (in FISH analyses)

MAL: Maximum Array Length (observed in MinIon reads of *L. migratoria*)

MST: Minimum Spanning Tree

Ode: *Oedaleus decorus*

OSF: Orthologous Superfamily

RL: Repeat Landscape

RPS: Relative peak size

RUL: Repeat Unit Length

satDNA: satellite DNA

SF: Superfamily

TSI: Tandem Structure Index

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

VIF: Variance inflation factors

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not aplicable

Availability of data and materials

The Illumina libraries used for this article are available in the Sequence Read Archive (SRA) with accession numbers SRR9649806 and SRR2911427. Main data generated or analyzed during this study are included in this published article and its supplementary information files. The remaining datasets can be requested to the corresponding author.

Competing interests

The authors declare no competing interests.

Funding

FJRR was also supported by a postdoctoral fellowship from Sven och Lilly Lawskis fond (Sweden) and a Marie Skłodowska-Curie Individual Fellowship (grant agreement 875732, European Union).

Authors' contributions

Conceptualization: JPMC, JC, MDLL, MMP, FP, MAGR, FJRR; experimental design: JPMC, JC, MDLL, MMP, FP, MAGR, FJRR; sampling: JPMC and JC; cytogenetic analyses: JPMC, JC, MDLL; data analysis: JPMC, MMP, MAGR, FJRR. All authors read and approved the manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

Arnason U, Grettarsdottir S, Widegren B. Mysticete (baleen whale) relationships based upon the sequence of the common cetacean DNA satellite. Mol. Biol. Evol. 1992, 9, 1018–1028.

Blount ZD, Lenski RE, Losos JB. Contingency and determinism in evolution: Replaying life's tape. Science. 2018;362(6415). https://doi.org/10.1126/science.aam5979

Brajković J, Feliciello I, Bruvo-Mađarić B, Ugarković Đ. Satellite DNA-like elements associated with genes within euchromatin of the beetle Tribolium castaneum. G3: Genes, Genomes, Genetics. 2012;2:931–941.

Brukner I, Sanchez R, Suck D, Pongor S. Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. The EMBO journal. 1995;14(8):1812–1818.

Cabrero J, Bakkali M, Bugrov A, Warchalowska-Sliwa E, López-León MD, Perfectti F, Camacho JPM. Multiregional origin of B chromosomes in the grasshopper *Eyprepocnemis plorans*. Chromosoma. 2003;112(4):207–211. https://doi.org/10.1007/s00412-003-0264-2

Cafasso D, Chinali G. An ancient satellite DNA has maintained repetitive units of the original structure in most species of the living fossil plant genus *Zamia*. Genome 2014;57:125–135.

Charlesworth B, Langley CH, Stephan W. The evolution of restricted recombination and the accumulation of repeated DNA sequences. Genetics. 1986;112(4):947–962.

Cohen S, Agmon N, Yacobi K, Mislovati M, Segal D. Evidence for rolling circle replication of tandem genes in Drosophila. Nucleic Acids Research. 2005;33(14):4519–4526. https://doi.org/10.1093/nar/gki764

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Cohen S, Agmon N, Sobol O, Segal D. Extrachromosomal circles of satellite repeats and 5S ribosomal DNA in human cells. Mobile DNA. 2010;1:11. https://doi.org/10.1186/1759-8753-1-11

de la Herrán R, Fontana F, Lanfredi M, Congiu L, Leis M, Rossi R, Ruiz Rejón C, Ruiz Rejón M, Garrido-Ramos MA.. Slow rates of evolution and sequence homogenization in an ancient satellite DNA family of sturgeons. Molecular Biology and Evolution. 2001a;18(3):432–436.

de La Herrán R, Ruiz Rejón C, Ruiz Rejón M, Garrido-Ramos MA. The molecular phylogeny of the Sparidae (Pisces, Perciformes) based on two satellite DNA families. Heredity. 2001b;87(6):691–697.

Djupedal I, Kos-Braun IC, Mosher RA, Söderholm N, Simmer F, Hardcastle TJ, Fender A, Heidrich N, Kagansky A, Bayne E, et al. Analysis of small RNA in fission yeast; centromeric siRNAs are potentially generated through a structured RNA. EMBO J. 2009;28:3832–3844.

Dover G. Molecular drive: a cohesive mode of species evolution. Nature. 1982a;299:111–117.

Dover G. A molecular drive through evolution. Bioscience. 1982b;32(6):526–33.

Drummond AJ, Ashton B, Buxton S, Cheung M, Cooper A, Heled J, Kearse M, Moir R, Stones-Havas S, Sturrock S, Thierer T, Wilson A. Geneious v. 4.8. Auckland, New Zealand: Biomatters Ltd. 2010.

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research. 2004;32(5):1792–1797.

Eirín-López JM, Rebordinos L, Rooney AP, Rozas J. The birth- and-death evolution of multigene families revisited. In: Garrido-Ramos MA, editor. Repetitive DNA. Basel: S. Karger AG; 2012. p. 170–196.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Escudeiro A, Adega F, Robinson TJ, Heslop-Harrison JS, Chaves R. Conservation, divergence and functions of centromeric satellite DNA families in the Bovidae. Genome Biology and Evolution, 2019;11(4):1152–1165. https://doi.org/10.1093/gbe/evz061

Excoffier L, Lischer HE. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and windows. Mol Ecol Resour. 2010;10:564–567. doi:10.1111/j.1755- 0998.2010.02847.x

Fachinetti D, Han JS, McMahon MA, Ly P, Abdullah A, Wong AJ, Cleveland DW. DNA sequence-specific binding of CENP-B enhances the fidelity of human centromere function. Dev. Cell. 2015;33:314–327.

Fry K, Salser W. Nucleotide sequences of HS-α satellite DNA from kangaroo rat dipodomys ordii and characterization of similar sequences in other rodents. Cell. 1977;12(4):1069–1084. https://doi.org/10.1016/0092-8674(77)90170-2

Gabrielian A, Pongor S. Correlation of intrinsic DNA curvature with DNA property periodicity. FEBS letters. 1996;393(1):65–68.

Gabrielian A, Simoncsits A, Pongor S. Distribution of bending propensity in DNA sequences. FEBS letters. 1996;393(1):124–130.

Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J. 1986;292:746–750.

Garrido-Ramos MA, Jamilena M, Lozano R, Ruiz Rejón C, Ruiz Rejón M. The EcoRI centromeric satellite DNA of the Sparidae family (Pisces, Perciformes) contains a sequence motive common to other vertebrate centromeric satellite DNAs. Cytogenet. Cell. Genet. 1995;71:345–351.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Garrido-Ramos MA, de la Herran R, Jamilena M, Lozano R, Ruiz Rejón C, Ruiz Rejón M. Evolution of centromeric satellite-DNA and its use in phylogenetic studies of the Sparidae family (Pisces, Perciformes). Mol. Phyl. Evol. 1999;12:200–204.

Garrido-Ramos MA. Satellite DNA in Plants: More than Just Rubbish. Cytogenetic and Genome Research. 2015;146(2):153–170. https://doi.org/10.1159/000437008

Garrido-Ramos MA. Satellite DNA: An evolving topic. Genes. 2017;8(9):230. https://doi.org/10.3390/genes8090230

Goodsell DS, Dickerson RE. Bending and curvature calculations in B-DNA. Nucleic acids research. 1994;22(24):5497–5503.

Gould SJ. Wonderful life: the Burgess Shale and the nature of history. Norton, New York; 1989.

Haaf T, Mater AG, Wienberg J, Ward DC. Presence and abundance of CENP-B box sequences in great ape subsets of primate-specific alpha-satellite DNA. J. Mol. Evol. 1995;41:487–491.

Hall SE, Kettler G, Preuss D. Centromere satellites from Arabidopsis populations: maintenance of conserved and variable domains. Genome research. 2003;13(2):195–205.

Harding RM, Boyce AJ, Clegg JB. The evolution of tandemly repetitive DNA: recombination rules. Genetics. 1992;132(3):847–859.

Ho J, Tumkaya T, Aryal S, Choi H, Claridge-Chang A. Moving beyond P values: data analysis with estimation graphics. Nature Methods. 2019;16(7):565–566. https://doi.org/10.1038/s41592-019-0470-3

Iwata A, Tek AL, Richard MM, Abernathy B, Fonsêca A, Schmutz J, Chen NWG, Thareau V, Magdelenat G, Li Y, Murata M, Pedrosa-Harand A, Geffroy V, Nagaki

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

K, Jackson SA. Identification and characterization of functional centromeres of the common bean. The Plant Journal. 2013;76(1):47–60.

Kasinathan S, Henikoff S. Non-B-form DNA is enriched at centromeres. Molecular Biology and Evolution. 2018;35(4):949–962. https://doi.org/10.1093/molbev/msy010

Kimura M, Ohta T. Population genetics of multigene family with special reference to decrease of genetic correlation with distance between gene members on a chromosome. Proc Nat Acad Sci USA. 1979;76(8):4001–4005.

Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of molecular evolution. 1980;16(2):111–120.

Kit S. Equilibrium sedimentation in density gradients of DNA preparations from animal tissues. Journal of Molecular Biology. 1961;3(6):711–716. https://doi.org/10.1016/S0022-2836(61)80075-2

Koch J. Neocentromeres and alpha satellite: a proposed structural code for functional human centromere DNA. Human molecular genetics. 2000;9(2):149–154.

Kuhn GCS, Küttler H, Moreira-Filho O, Heslop-Harrison JS. The 1.688 repetitive DNA of drosophila: Concerted evolution at different genomic scales and association with genes. Molecular Biology and Evolution. 2012;29(1):7–11. https://doi.org/10.1093/molbev/msr173

Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009;25(11):1451–1452.

Lorite P, Muñoz-López M, Carrillo JA, Sanllorente O, Vela J, Mora P, Tinaut A, Torres MI, Palomeque T. Concerted evolution, a slow process for ant satellite DNA: study

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of the satellite DNA in the *Aphaenogaster* genus (Hymenoptera, Formicidae). Organisms Diversity & Evolution. 2017;17(3):595–606.

Luchetti A, Cesari M, Carrara G, Cavicchi S, Passamonti M, Scali V, Mantovani B. Unisexuality and molecular drive: Bag320 sequence diversity in *Bacillus* taxa (Insecta Phasmatodea). Journal of molecular evolution. 2003;56(5):587–596.

Luchetti A, Marini M, Mantovani B. Non-concerted evolution of the RET76 satellite DNA family in *Reticulitermes* taxa (Insecta, Isoptera). Genetica. 2006;128:123–132.

Lynch M. Statistical inference on the mechanisms of genome evolution. PLoS Genetics. 2011;7(6):1–4. https://doi.org/10.1371/journal.pgen.1001389

Lynch M, Bobay L-M, Catania F, Gout J-F, Rho M. The Repatterning of Eukaryotic Genomes by Random Genetic Drift. Annu Rev Genomics Hum Genet. 2011;12:347–366. https://doi.org/10.1146/annurev-genom-082410-101412

Masumoto H, Masukata H, Muro Y, Nozaki N, Okazaki T. A human centromere antigen (CENP-B) interacts with a short specific sequence in alphoid DNA, a human centromeric satellite. J. Cell Biol. 1989:109:1963–1973.

Masumoto H, Nakano M, Ohzeki J. The role of CENP-B and alpha-satellite DNA: De novo assembly and epigenetic maintenance of human centromeres. Chromosome Res. 2004;12:543–556.

Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, Sebra R, Peluso P, Eid J, Rank D, Garcia JF. Genome Biol. 2013;14:R10.

Meštrović N, Mravinac B, Pavlek M, Vojvoda-Zeljko T, Šatović E, Plohl M. Structural and functional liaisons between transposable elements and satellite DNAs. Chromosome Research. 2015;23(3):583–596. https://doi.org/10.1007/s10577-015-9483-7

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Mravinac B, Plohl M, Meštrović N, Ugarković Đ. Sequence of PRAT satellite DNA "frozen" in some Coleopteran species. J. Mol. Evol. 2002;54:774–783.

Mravinac B, Plohl M, Ugarković Đ. Preservation and high sequence conservation of satellite DNAs suggest functional constraints. J. Mol. Evol. 2005;61:542–550.

Muro Y, Masumoto H, Yoda K, Nozaki N, Ohashi M, Okazaki T. Centromere protein B assembles human centromeric alpha-satellite DNA at the 17-bp sequence, CENP-B box. J Cell Biol. 1992;116:585–596.

Navajas-Perez R, de la Herrán R, Jamilena M, Lozano R, Ruiz Rejon C, Ruiz Rejon M, Garrido-Ramos MA. Reduced rates of sequence evolution of Y-linked satellite DNA in *Rumex* (Polygonaceae). Journal of molecular evolution. 2005;60(3):391–399.

Navajas-Pérez R, Schwarzacher T, de la Herrán R, Ruiz Rejón C, Ruiz Rejón M, Garrido-Ramos MA. The origin and evolution of the variability in a Y-specific satellite-DNA of *Rumex acetosa* and its relatives. Gene. 2006;368:61–71.

Navajas-Pérez R, Quesada del Bosque ME, Garrido-Ramos MA. Effect of location, organization, and repeat-copy number in satellite-DNA evolution. Molecular Genetics and Genomics. 2009;282(4):395–406.

Nei M, Rooney AP. Concerted and birth-and-death evolution of multigene families. Annu Rev Genet. 2005;39:121–152.

Neumann P, Navrátilová A, Schroeder-Reiter E, Koblížková A, Steinbauerová V, Chocholová E, Novák P, Wanner G, Macas J. Stretching the rules: monocentric chromosomes with multiple centromere domains. PLoS Genet. 2012;8(6):e1002777.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Novák P, Neumann P, Pech J, Steinhaisl J, Macas J. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics. 2013;29:792–793.

Ohta T. Genetic variation in small multigene families. Genetical Research. 1981;37(2):133–149. https://doi.org/10.1017/S0016672300020115

Ohta T. On the evolution of multigene families. Theoretical Population Biology. 1983;23(2);216–240. https://doi.org/10.1016/0040-5809(83)90015-1

Ohta T, Kimura M. Some calculations on the amount of selfish DNA. Proc Natl Acad Sci USA. 1981;78:1129–1132. doi: 10.1073/pnas.78.2.1129.

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 Research. 2015;22(6):387–401. https://doi.org/10.1093/dnares/dsv021

Pezer Ž, Brajković J, Feliciello I, Ugarković Đ. Satellite DNA-mediated effects on genome regulation. Repetitive DNA. 2012;7:153–169.

Plohl M, Meštrovic N, Mravinac B. Satellite DNA evolution.In: Garrido-Ramos MA, editor. Repetitive DNA. Basel: S. Karger AG; 2012. p. 126–152.

Quesada del Bosque ME, López-Flores I, Suárez-Santiago VN, Garrido-Ramos MA. Differential spreading of Hin fI satellite DNA variants during radiation in Centaureinae. Annals of botany. 2013;112(9):1793–1802.

Quesada del Bosque ME, López-Flores I, Suárez-Santiago VN, Garrido-Ramos MA. Satellite-DNA diversification and the evolution of major lineages in Cardueae (Carduoideae Asteraceae). Journal of plant research. 2014;127(5):575–583.

Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. Trends in genetics. 2000;16(6):276–277.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Robles F, de la Herrán R, Ludwig A, Ruiz Rejón C, Ruiz Rejón M, Garrido-Ramos MA. Evolution of ancient satellite DNAs in sturgeon genomes. Gene. 2004;338:133–142.

Rousset F, Raymond M. Testing heterozygote excess and deficiency. Genetics 1995;140:1413–1419.

Ruiz-Ruano FJ, López-León MD, Cabrero J, Camacho JPM. High-throughput analysis of the satellitome illuminates satellite DNA evolution. Scientific Reports. 2016;6:28333. https://doi.org/10.1038/srep28333

Ruiz-Ruano FJ, Cabrero J, López-León MD, Camacho JPM. Satellite DNA content illuminates the ancestry of a supernumerary (B) chromosome. Chromosoma. 2017;126:487–500. https://doi.org/10.1007/s00412-016-0611-8

Ruiz-Ruano FJ, Castillo-Martínez J, Cabrero J, Gómez R, Camacho JPM, López-León MD. High-throughput analysis of satellite DNA in the grasshopper *Pyrgomorpha conica* reveals abundance of homologous and heterologous higher-order repeats. Chromosoma. 2018a;127(3):323–340. https://doi.org/10.1007/s00412-018-0666-9

Ruiz-Ruano FJ, Cabrero J, López-León MD, Sánchez A, Camacho JPM. Quantitative sequence characterization for repetitive DNA content in the supernumerary chromosome of the migratory locust. Chromosoma. 2018b;127(1):45–57.

SantaLucia Jr J. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proceedings of the National Academy of Sciences. 1998;95(4):1460–1465.

Šatović E, Plohl M. Tandem Repeat-Containing MITEs in the Clam *Donax trunculus*. Genome Biology and Evolution. 2013;5(12):2549–2559. https://doi.org/10.1093/gbe/evt202

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Šatović E, Vojvoda Zeljko T, Luchetti A, Mantovani B, Plohl M. Adjacent sequences disclose potential for intra-genomic dispersal of satellite DNA repeats and suggest a complex network with transposable elements. BMC Genomics. 2016;17(1):997. https://doi.org/10.1186/s12864-016-3347-1

Schmieder R, Edwards R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS One. 2011;6:e17288.

Schueler MG, Swanson W, Thomas PJ. NISC Comparative Sequencing Program & Green, E.D. Adaptive evolution of foundation kinetochore proteins in primates. Mol. Biol. Evol. 2010;27:1585–1597.

Shang WH, Hori T, Toyoda A, Kato J, Popendorf K, Sakakibara Y, Fujiyama A, Fukagawa T. Chickens possess centromeres with both extended tandem repeats and short non-tandem-repetitive sequences. Genome research. 2010;20(9):1219–1228.

Smit AFA, Hubley R, Green P (2013) RepeatMasker Open-4.0. http://www.repeatmasker.org

Smith G P. Evolution of repeated DNA sequences by unequal crossover. Science. 1976;191(4227):528–535. https://doi.org/10.1126/science.1251186

Song H, Amédégnato C, Cigliano MM, Desutter-Grandcolas L, Heads SW, Huang Y, Otte D, Whiting MF. 300 million years of diversification: elucidating the patterns of orthopteran evolution based on comprehensive taxon and gene sampling. Cladistics. 2015;31:621–651. https://doi.org/https://doi.org/10.1111/cla.12116

Stephan W. Recombination and the evolution of satellite DNA. Genet. Res. 1986;47:167–174.

Stephan W. Quantitative variation and chromosomal location of satellite DNAs. Genet. Res. 1987;50(1):41–52.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Stephan W. Tandem-repetitive non coding DNA: forms and forces. Molecular Biology and Evolution. 1989;6(2):198–212. https://doi.org/10.1093/oxfordjournals.molbev.a040542

Stephan W, Cho S. Possible role of natural selection in the formation of tandem-repetitive noncoding DNA. Genetics. 1994;136:333–341.

Suárez-Santiago VN, Blanca G, Ruiz-Rejón M, Garrido-Ramos MA. Satellite-DNA evolutionary patterns under a complex evolutionary scenario: The case of *Acrolophus* subgroup (*Centaurea* L., Compositae) from the western Mediterranean. Gene. 2007;404(1-2):80–92.

Talbert PB, Henikoff S. Transcribing centromeres: noncoding RNAs and kinetochore assembly. Trends in Genetics. 2018;34(8):587–599.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. Molecular biology and evolution. 2013;30(12):2725–2729.

Teacher AGF, Griffiths DJ. HapStar: automated haplotype network layout and visualization. Mol. Ecol. Resour. 2011;11:151–153. doi: 10.1111/j.1755-0998.2010.02890.x

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic acids research. 1997;25(24):4876–4882.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. Primer3—new capabilities and interfaces. Nucleic acids research. 2012;40(15):e115

Vlahovicek K, Kajan L, Pongor S. DNA analysis servers: plot. it, bend. it, model. it and IS. Nucleic Acids Research. 2003;31(13):3686–3687.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Walsh JB. Persistence of tandem arrays: implications for satellite and simple-sequence DNAs. Genetics. 1987;115(3):553–567.

Waring M, Britten RJ. Nucleotide sequence repetition: A rapidly reassociating fraction of mouse DNA. Science. 1966;154(3750):791–794. https://doi.org/10.1126/science.154.3750.791

Willard HF, Waye JS. Chromosome-specific subsets of human alpha satellite DNA: analysis of sequence divergence within and between chromosomal subsets and evidence for an ancestral pentameric repeat. J Mol Evol. 1987;25:207–214. https://doi.org/10. 1007/BF02100014

Zhang H, Koblížková A, Wang K, Gong Z, Oliveira L, Torres GA, Wu YF, Zhang W, Novák P, Buell CR, Macas J, Jiang J. Boom-bust turnovers of megabase-sized centromeric DNA in *Solanum* species: rapid evolution of DNA sequences associated with centromeres. The Plant Cell. 2014;26(4):1436–1447.

Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic acids research. 2003;31(13):3406–3415.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figures and Supplementary Figure

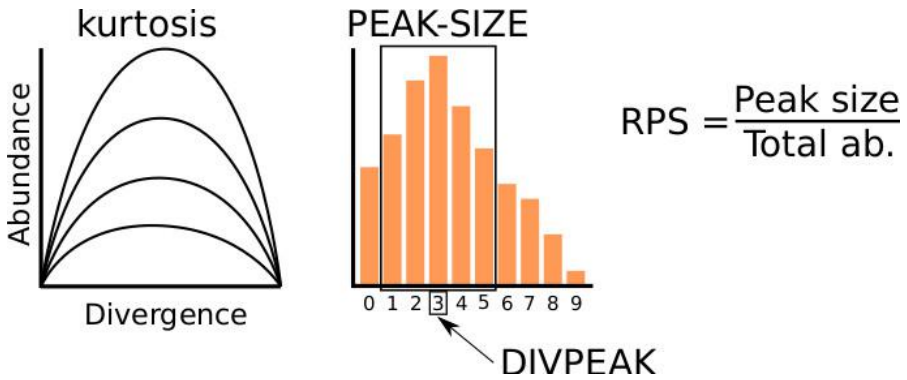

Figure 1. Definition of satDNA parameters in respect to abundance and divergence. The distribution of the abundances of groups of sequences differing by 1% divergence constitutes a repeat landscape (RL). It may be seen as a curve (left) or an histogram (right). In addition of variation in kurtosis, represented by several curves on the left, three properties of satDNA can be defined on RLs: DIVPEAK is the divergence class showing the highest abundance (3% in the histogram); PEAK-SIZE is the sum of the abundances of the five classes included around DIVPEAK, thus constituting the sum of all sequences differing by less than 5%, thus coinciding with our definition of satDNA subfamily; RPS is the relative peak size and represents the fraction of abundance which is included in the 5% amplification peak.

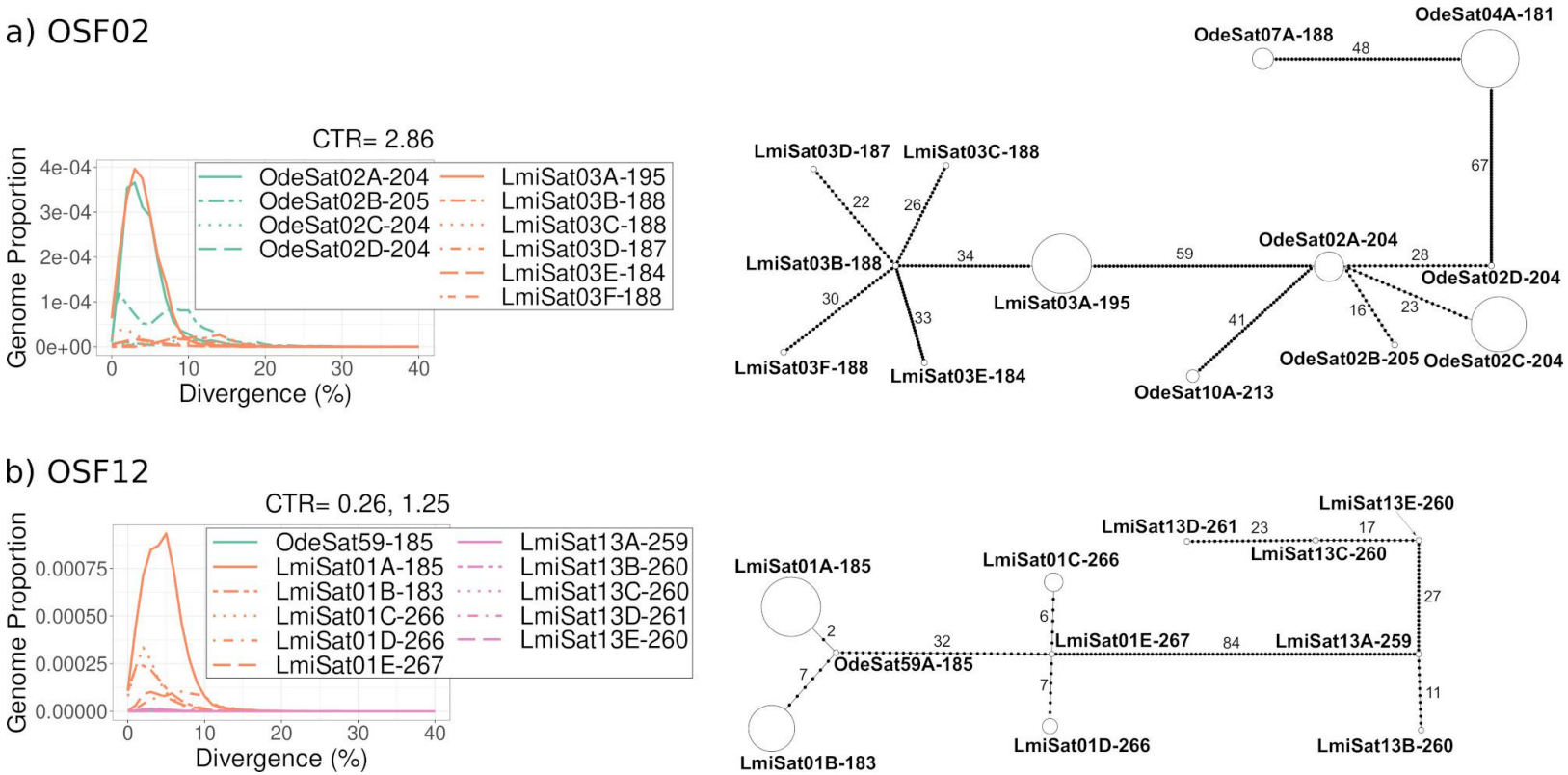

Figure 2. Repeat landscape (RL) and minimum spanning tree (MST) of two orthologous superfamilies of satellite DNA in *O. decorus* and *L. migratoria* (OSF02 and OSF12). a) OSF02 showed the highest consensus turnover rate (CTR= 2.86) found among the 20 values estimated between orthologous pairs of families in both species. Note that OSF02 showed large amplification peaks in both species (green curve in *O. decorus* and red curve in *L. migratoria*) and that the MST showed complete separation of OdeSat02 and LmiSat03 sequences. b) OSF12 showed the lowest CTR estimate (0.26 between OdeSat59 and LmiSat01) and the MST (on the right) reveals that the consensus DNA sequences of these two satDNA families showed only two differences. Also note in the RL (on the left) that the OdeSat59 curve is very close to zero, as this is the satDNA family in *O. decorus* showing the lowest abundance, indicating that OSF12 is represented in this species as relict remains which, by chance, almost coincide in consensus sequence with the most abundant subfamily in *L. migratoria* (LmiSat01A), thus evidencing extreme incomplete lineage sorting (see other cases in Fig. S1).

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

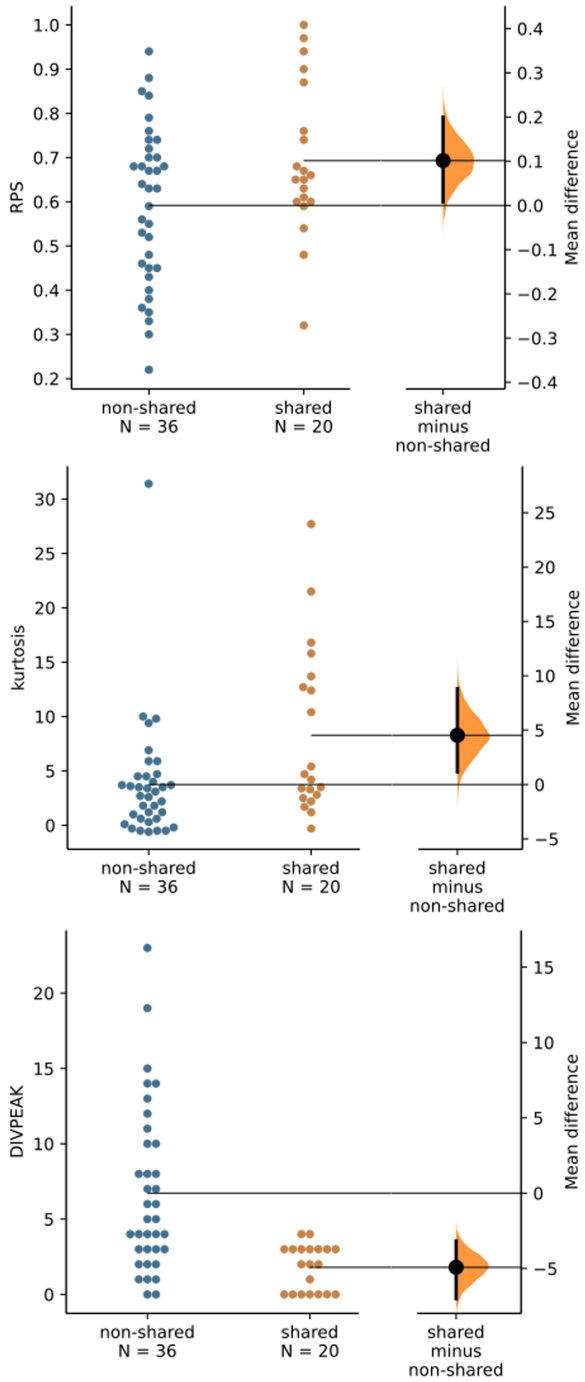

Figure 3. Gardner-Altman plots comparing RPS, kurtosis and DIVPEAK between the *L migratoria* satDNA families being shared or non-shared with *O. decorus*. Note that shared satDNAs showed higher homogenization (higher RPS and kurtosis) and lower degeneration (5% effect size for mean difference in DIVPEAK) than non-shared ones, suggesting most recent amplification of the shared ones.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

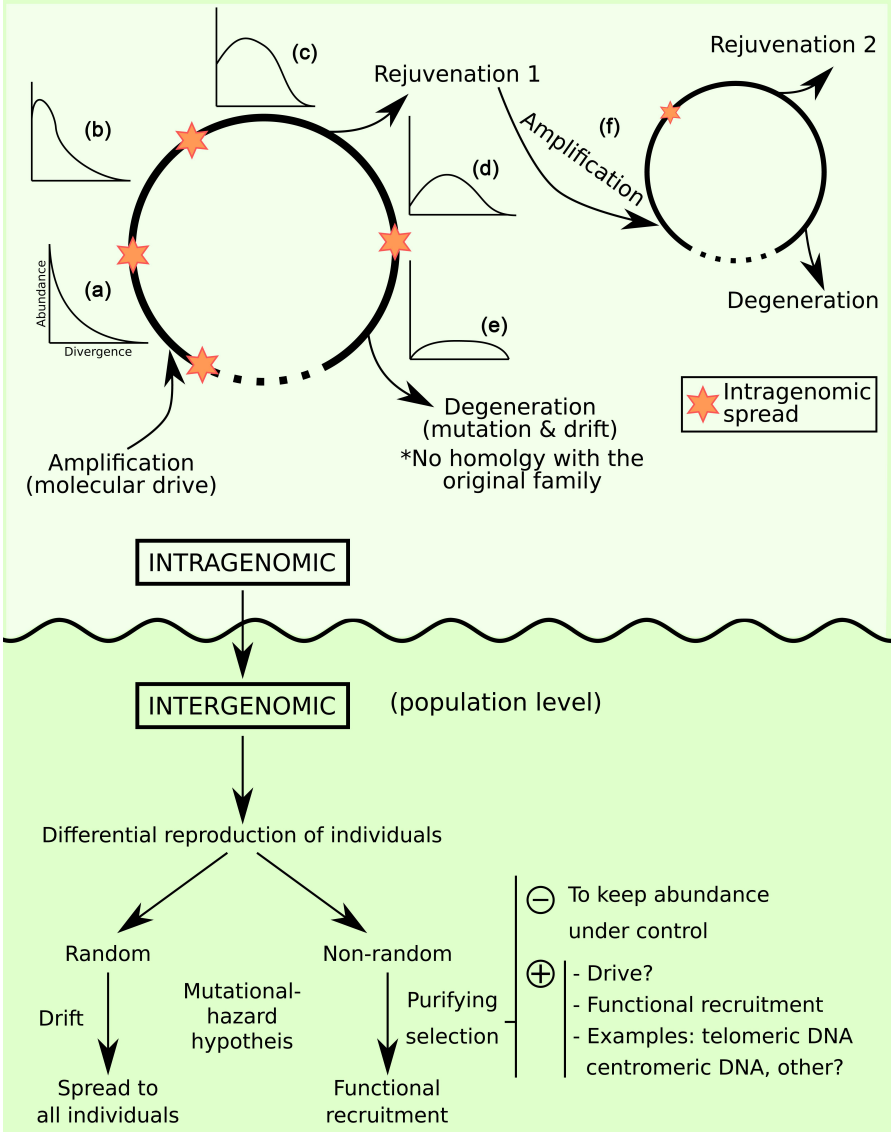

Figure 4. A model of satDNA evolution. We consider that evolutionary events are rather different at intra- and inter-genomic levels. At intragenomic level, tandem duplication yields many copies of a non-coding sequence which will essentially show the same sequence, thus displaying RLs sharply leptokurtic (a). As time goes by, point mutation increases divergence among the amplified sequences and the curve progressively is flattened (b-e) with increasing DIVPEAK. At any moment of this first amplification-degeneration cycle, another sequence undergoes amplification (f) and begins a cycle of rejuvenation-degeneration, and so on. In parallel, an intragenomic spread of the satDNA can occur at higher or lesser extent. A conceivable exit of these cycles is satDNA degeneration, when homology with the original sequence has been lost. At intergenomic level, individual reproduction will mark the destiny of the different satDNA sequences in populations. When reproduction is differential, albeit random (drift) or non-random (selection), some sequences may become prevalent above others. At this respect, the mutational-hazard hypothesis is applicable to explain the limits to purifying selection in some species showing extremely high abundance of satDNA. Finally, we cannot rule out that, in some case, drive could help satDNA to prosper and, even that positive selection may recruit satDNA for important functions, such as telomeric or centromeric functions.

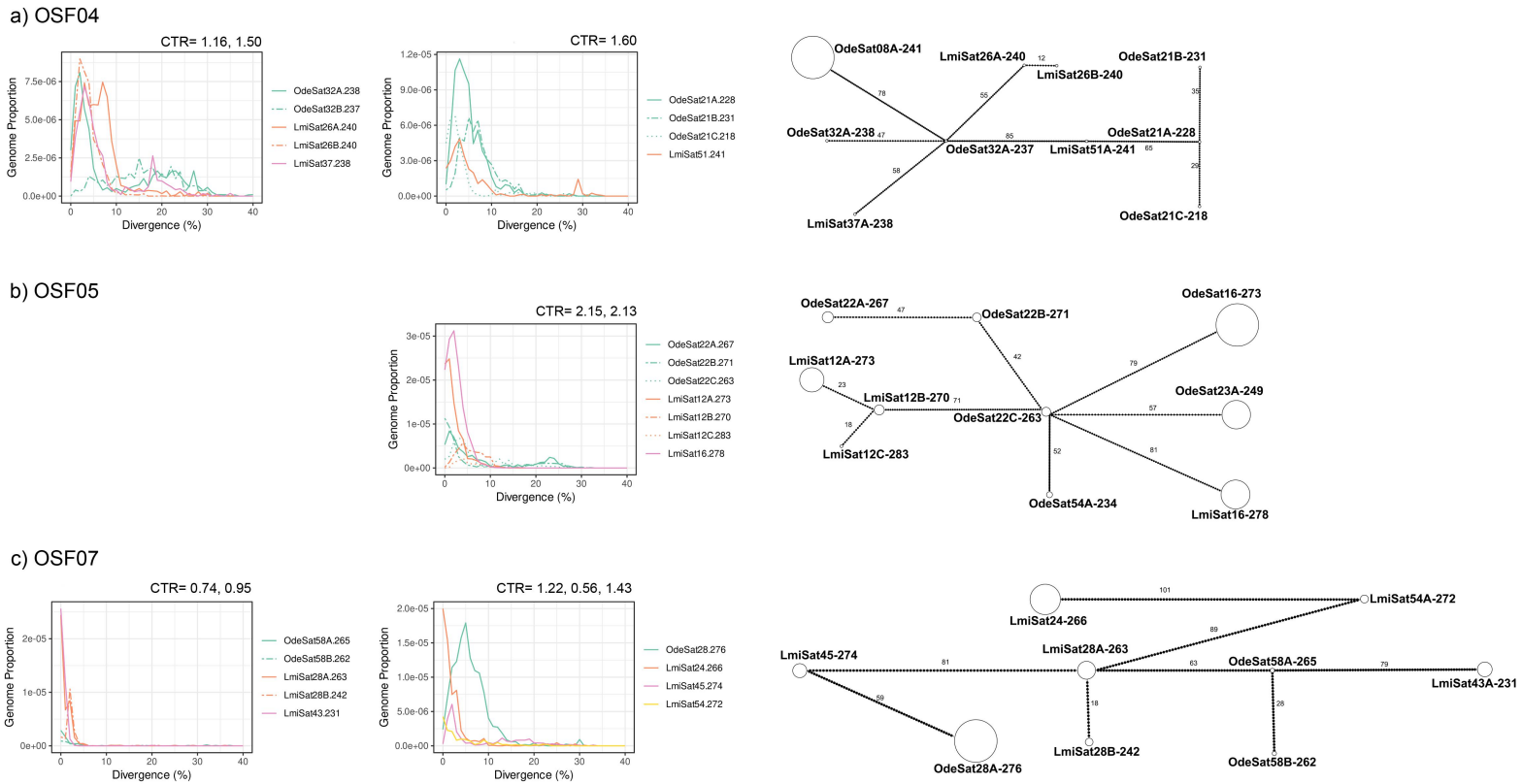

Figure S1. Repeat landscape (RL) and minimum spanning tree (MST) of three orthologous superfamilies of satellite DNA in *O. decorus* and *L. migratoria* (OSF04, OSF05 and OSF07). a) RLs showed that OSF04 showed large peaks of amplification in both species but CTR values ranged between 1.16 and 1.6, presumably due to the incomplete library sorting (ILibS) evidenced by the MST (note how OdeSat32A and LmiSat51A connect with both species' sequences). b) OSF05 showed high CTR values, large amplification peaks in both species and ILibS for only OdeSat22C, which was the only sequence connected with sequences from both species. c) OSF07 showed the lowest CTR values and showed very small amplification peaks for OdeSat58 (green curves in the RL on the left) and higher ILibS, with three sequences being connected with both species' sequences (LmiSat45-274, LmiSat28A-263 and OdeSat58A-265).

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450629; this version posted July 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Tables

Table 1. Comparison of satellitome characteristics between *O. decorus* and *L. migratoria* (Southern Lineage), by means of estimation graphics using DABEST (Ho et al. 2019). 95% CI= Confidence interval. RUL= Repeat unit length. * means that 95% CI does not include the zero value.

Comparison	Item	Mean (SE)		Effect size			
		***O. decorus* (N= 58)**	***L. migratoria* (N=56)**	**Unpaired mean difference**	**CI_low**	**CI_high**	**Includes zero?**
All satDNAs	RUL	201.5 (13.6)	152.7 (14)	48.8	12.1	86.6	*
	A+T (%)	55.7 (1.2)	54.4 (1.1)	1.27	-1.81	4.38	
	Abundance (%)	0.044 (0.013)	0.038 (0.019)	0.0055	-0.0557	0.0415	
	Divergence	7.19 (0.56)	7.09 (0.61)	0.093	-1.55	1.75	
		***O. decorus* (N= 21)**	***L. migratoria* (N= 20)**				
Shared satDNAs	RUL	212.8 (12.6)	216.5 (14.1)	-3.69	-39.4	33.3	
	A+T (%)	58.3 (1.1)	58.0 (1.1)	0.333	-2.8	3.27	
	Abundance (%)	0.071 (0.033)	0.091 (0.052)	-0.0196	-0.171	0.0715	
	Divergence	8.08 (1.22)	4.90 (0.50)	3.18	1.19	6.34	*
		***O. decorus* (N= 37)**	***L. migratoria* (N= 36)**				
Non-shared satDNAs	RUL	195.1 (20.2)	117.2 (17.8)	77.9	26.7	129	*
	A+T (%)	54.2 (1.7)	52.5 (1.6)	1.76	-2.75	6.21	
	Abundance (%)	0.028 (0.01)	0.009 (0.002)	0.019	0.00635	0.0496	*
	Divergence	6.68 (0.53)	8.31 (0.84)	-1.63	-3.64	0.244	
		Shared (N= 21)	**Non-shared (N= 37)**				
O. decorus	RUL	212.8 (12.6)	195.1 (20.2)	17.7	-34.4	58.3	
	A+T (%)	58.3 (1.1)	54.2 (1.7)	4.11	0.299	8.19	*
	Abundance (%)	0.071 (0.033)	0.028 (0.01)	0.0434	-0.00243	0.139	
	Divergence	8.08 (1.22)	6.68 (0.53)	1.4	-0.699	4.63	
		Shared (N= 20)	**Non-shared (N= 36)**				
L. migratoria	RUL	216.5 (14.1)	117.2 (17.8)	99.3	50	139	*
	A+T (%)	58.0 (1.1)	52.5 (1.6)	5.45	1.95	9.43	*
	Abundance (%)	0.091 (0.052)	0.009 (0.002)	0.082	0.018	0.261	*
	Divergence	4.90 (0.50)	8.31 (0.84)	-3.41	-5.42	-1.59	*

1446

Table 2. Characteristics of the orthologous satDNA families analyzed in *O. decorus* (14) and *L. migratoria* (20). Each row includes one Ode and one Lmi satDNA families showing homology between them. Note that some Ode families showed homology with two or three Lmi ones. OSF= Orthologous superfamily, sf= number of subfamilies, FISH= FISH pattern (B= banded, NS= no signal), abun= abundance (% of the genome), RPS= Relative peak size, DP= DIVPEAK, MAL= Maximum array length observed in MinIon reads of L. migratoria, CEI= Concerted evolution index (L= *L. migratoria*, O= *O. decorus*), Intid= Interspecific sequence identity (%), Intdiv= Interspecific divergence, CTR= Consensus turnover rate, ILibS= Incomplete library sorting. Negative CEI values and Int_id>95% are remarked in bold type letter. See Table S4 to complete data with repeat unit length, A+T content, divergence (%), peak size, kurtosis of the repeat landscape, tandem structure index and Gibbs free energy of the secondary structure.

	O. decorus						***Locusta migratoria***							**Interspecific comparisons**					
OSF	**Name**	**sf**	**FISH**	**abun**	**RPS**	**DP**	**Name**	**sf**	**FISH**	**abun**	**RPS**	**DP**	**MAL**	**CEI_O**	**CEI_L**	**Int_id**	**Int_div**	**CTR**	**ILibS**
1	OdeSat01-287	1	B	6.2E-03	87%	1	LmiSat09-181	5	B	3.0E-04	65%	0	4417	88.4	85.6	68.9	90.8	1.990	0.30
2	OdeSat02-204	4	B	3.3E-03	51%	2	LmiSat03-195	6	B	3.0E-03	63%	3	13447	124.5	125.1	60.6	130.4	2.858	0
3	OdeSat17-176	1	NS	2.0E-04	29%	27	LmiSat02-176	1	B	3.6E-03	68%	4	20180	**-24.6**	**-5.1**	**99.4**	0.6	0.013	1.00
4	OdeSat21-228	3	NS	1.5E-04	58%	3	LmiSat51-241	1	B	2.9E-05	61%	3	1708	67.0	66.5	71.8	72.8	1.596	0.44
4	OdeSat32-238	2	B	8.5E-05	36%	2	LmiSat26-240	2	B	1.0E-04	60%	3	1455	40.5	47.8	77.7	53.1	1.164	0.59
4							LmiSat37-238	1	B	4.6E-05	59%	3	2454	54.4	59.5	75.6	67	1.469	0.49
5	OdeSat22-267	3	B	1.4E-04	59%	1	LmiSat12-273	3	B	1.3E-04	74%	1	2948	90.6	94.8	75	98.1	2.150	0.25
5							LmiSat16-278	1	B	1.4E-04	87%	2	1965	89.5	94.6	72.6	97	2.126	0.26
6	OdeSat26-180	1	B	1.3E-04	88%	2	LmiSat41-180	1	B	5.1E-05	94%	3	515	29.2	28.2	74.4	31.7	0.695	0.76
7	OdeSat28-276	1	B	1.2E-04	56%	5	LmiSat24-266	1	NS	5.9E-05	90%	0	1378	49.4	53.4	67.9	55.8	1.223	0.57
7							LmiSat45-274	1	B	2.5E-05	54%	2	945	19.0	16.8	79.7	25.4	0.557	0.81
7							LmiSat54-272	1	B	1.6E-05	65%	0	2073	58.7	60.2	66.3	65.1	1.427	0.50
7	OdeSat58-265	2	NS	9.5E-06	88%	0	LmiSat28-263	2	B	6.0E-05	97%	0	2821	30.1	32.4	77.5	33.9	0.743	0.74
7							LmiSat43-231	1	B	3.9E-05	100%	0		39.3	42.7	69.3	43.1	0.945	0.67
8	OdeSat39-185	2	NS	6.8E-05	67%	4	LmiSat06-185	4	B	4.9E-04	66%	3	19168	14.9	16.1	84.3	21	0.460	0.84
9	OdeSat41-75	1	NS	6.1E-05	29%	18	LmiSat27-57	1	NS	5.4E-05	32%	0	712	**-2.4**	7.1	92.7	16.2	0.355	0.88
10	OdeSat56-249	1	NS	2.0E-05	93%	0	LmiSat32-261	1	B	3.9E-05	60%	0	1489	31.5	26.4	77.2	32.9	0.721	0.75
11	OdeSat57-75	1	NS	1.4E-05	40%	4	LmiSat17-75	1	B	1.2E-04	48%	2	3194	**-1.3**	2.7	92	8.5	0.186	0.93
12	OdeSat59-185	1	NS	5.8E-06	36%	3	LmiSat01-185	5	B	9.8E-03	46%	3	17619	**-0.9**	7.2	**98.9**	11.8	0.259	0.91
12							LmiSat13-259	5	B	1.5E-04	76%	4	1379	44.1	52.3	63.3	56.8	1.245	0.56
															Mean	77.3	50.6	1.109	61%
															SD	11.1	34.7	0.76	27%
															CV	14%	69%	69%	44%

1447

1448

Table 3. Spearman rank correlation (rS) between satellitome characteristics in *Oedaleus decorus* (Ode) and *Locusta migratoria* (Lmi). Pb= Sequential Bonferroni correction. RUL= Repeat unit length, TSI= Tandem structure index. RPS= Relative peak size.

	Ode (N=58)				Lmi (N= 56)			
Ode	**rS**	**t(N-2)**	**P**	**Pb**	**rS**	**t(N-2)**	**P**	**Pb**
Divergence & RUL	-0.29	-2.23	3.0E-02	3.6E-01	-0.39	-3.07	3.4E-03	**4.7E-02**
Divergence & A+T (%)	-0.20	-1.51	1.4E-01	1.4E+00	-0.11	-0.83	4.1E-01	2.9E+00
Divergence & subfam	0.02	0.18	8.6E-01	8.6E-01	-0.04	-0.27	7.9E-01	7.9E-01
Divergence & TSI	-0.56	-5.11	4.0E-06	**6.0E-05**	-0.26	-1.94	5.7E-02	6.9E-01
Divergence & Abundance	0.03	0.19	8.5E-01	2.5E+00	-0.24	-1.83	7.2E-02	7.9E-01
Divergence & RPS	**-0.89**	-14.91	3.8E-21	**6.5E-20**	**-0.90**	-15.09	5.1E-21	**8.7E-20**
RPS & RUL	0.14	1.05	3.0E-01	2.4E+00	0.31	2.38	2.1E-02	2.7E-01
RPS & A+T (%)	0.10	0.76	4.5E-01	2.7E+00	0.05	0.39	7.0E-01	2.8E+00
RPS & subfam	-0.02	-0.19	8.5E-01	1.7E+00	0.04	0.28	7.8E-01	1.6E+00
RPS & TSI	0.58	5.32	1.9E-06	**3.0E-05**	0.21	1.59	1.2E-01	1.2E+00
RPS & Abundance	-0.14	-1.03	3.1E-01	2.2E+00	0.11	0.85	4.0E-01	3.2E+00
RPS & DIVPEAK	**-0.53**	-4.74	1.5E-05	**2.0E-04**	**-0.63**	-5.98	1.8E-07	**2.9E-06**
DIVPEAK & RUL	**-0.55**	-4.99	6.2E-06	**8.6E-05**	**-0.52**	-4.47	4.0E-05	**6.0E-04**
DIVPEAK & A+T (%)	-0.20	-1.55	1.3E-01	1.4E+00	-0.10	-0.75	4.6E-01	2.7E+00
DIVPEAK & subfam	-0.07	-0.50	6.2E-01	3.1E+00	0.05	0.33	7.4E-01	2.2E+00
DIVPEAK & TSI	-0.05	-0.40	6.9E-01	2.8E+00	0.07	0.53	6.0E-01	3.0E+00
DIVPEAK & Abundance	0.17	1.29	2.0E-01	1.8E+00	-0.13	-0.99	3.2E-01	2.9E+00

1449
1450
1451
1452
1453
1454
1455
1456
1457

1458

Table 4. Stepwise multiple regression of CTR, estimated from 14 orthologous pairs of satDNA families, on four satellitome features of *L. migratoria* (Lmi) and *O. decorus* (Ode). In each species, the independent variables employed were the number of subfamilies (subfam), the number of repeat units included in amplification peaks (peak_copies), the tandem structure index (TSI) and the homogenization index (RPS). Note that only three independent variables entered in the model, all of them corresponding to Ode, and only two (Ode_subfam and Ode_peak_copies) were associated with significant increases in explained variance in CTR (56.4% and 25.7%, respectively. The multiple correlation coefficients were 0.652 (SE= 0.13) and 0.466 (SE= 0.127), respectively. The Shapiro-Wilks test showed that the standardized residuals of this regression fitted a normal distribution (W= 0.966, P= 0.821). VIF= Variance inflation factors. Redundancy r^2 was performed between each independent item and the seven remaining, in order to calculate VIF as 1/(1-r^2).

Item	Redundancy r^2	VIF	Step	Multiple r	Multiple r^2	r^2 increase	F	P	Partial r
Lmi_subfam	0.668	3.01							
Lmi_peak_copies	0.262	1.36							
Lmi_TSI	0.331	1.50							
Lmi_RPS	0.236	1.31							
Ode_subfam	0.118	1.13	1	0.751	0.564	**0.564**	15.54	0.0020	0.845
Ode_peak_copies	0.067	1.07	2	0.906	0.822	**0.257**	15.89	0.0021	0.758
Ode_TSI	0.172	1.21	3	0.922	0.850	**0.028**	1.89	0.1995	0.398
Ode_RPS	0.565	2.30							

1459