1Too much too many: comparative analysis of morabine grasshopper genomes reveals highly 2abundant transposable elements and rapidly proliferating satellite DNA repeats

3

4Octavio M. Palacios-Gimenez^{1,2}, Julia Koelman¹, Marc Palmada Flores¹, Tessa M. Bradford^{3,4}, Karl
5K. Jones³, Steven J. B. Cooper^{3,4}, Takeshi Kawakami^{1,5*}, Alexander Suh^{1,2,6*}

6

7¹Department of Ecology and Genetics – Evolutionary Biology, Evolutionary Biology Centre,
8Uppsala University, SE-752 36 Uppsala, Sweden

9²Department of Organismal Biology – Systematic Biology, Evolutionary Biology Centre, Uppsala
10University, SE-752 36 Uppsala, Sweden

11³Evolutionary Biology Unit, South Australian Museum, Adelaide, SA 5000, Australia

12⁴School of Biological Sciences and Australian Centre for Evolutionary Biology and Biodiversity,
13The University of Adelaide, Adelaide, SA 5005, Australia

14⁵Embark Veterinary, Inc., Boston, Massachusetts, United States of America

15⁶School of Biological Sciences, University of East Anglia, Norwich Research Park, NR4 7TU,
16Norwich, United Kingdom

17

18*equal contributions

19

20**Correspondence:** Octavio M. Palacios-Gimenez, Alexander Suh and Takeshi Kawakami,
21Department of Organismal Biology – Systematic Biology, Evolutionary Biology Centre, Uppsala
22University, SE-752 36 Uppsala, Sweden

23

24**Emails:** octavio.palacios@ebc.uu.se (O.M.P.G.), alexander.suh@ebc.uu.se (A.S.),
25tkawakami@embarkvet.com (T.K.)

26Background: The repeatome, the collection of repetitive DNA sequences represented by
 27transposable elements (TEs) and tandemly repeated satellite DNA (satDNAs), is found in high
 28proportion in organisms across the tree of life. Grasshoppers have large genomes (average 9 Gb),
 29containing large amounts of repetitive DNA which has hampered progress in assembling reference
 30genomes. Here we combined linked-read genomics with transcriptomics to assemble, characterize,
 31and compare the structure of the repeatome and its contribution to genome evolution, in four
 32chromosomal races of the morabine grasshopper *Vandiemennella viatica* species complex.

33Results: We obtained linked-read genome assemblies of 2.73-3.27 Gb from estimated genome sizes
 34of 4.26-5.07 Gb DNA per haploid genome of the four chromosomal races of *V. viatica*. These
 35constitute the third largest insect genomes assembled so far (the largest being two locust
 36grasshoppers). Combining complementary annotation tools and manual curation, we found a large
 37diversity of TEs and satDNAs constituting 66 to 75 % per genome assembly. A comparison of
 38sequence divergence within the TE classes revealed massive accumulation of recent TEs in all four
 39races (314-463 Mb per assembly), indicating that their large genome size is likely due to similar
 40rates of TE accumulation across the four races. Transcriptome sequencing showed more biased TE
 41expression in reproductive tissues than somatic tissues, implying permissive transcription in
 42gametogenesis. Out of 129 satDNA families, 102 satDNA families were shared among the four
 43chromosomal races, which likely represent a repertoire of satDNA families in the ancestor of the *V.*
 44*viatica* chromosomal races. Notably, 50 of these shared satDNA families underwent differential
 45proliferation since the recent diversification of the *V. viatica* species complex.

46Conclusion: In-depth annotation of the repeatome in morabine grasshoppers provided new insights
 47into the genome evolution of Orthoptera. Our TEs analysis revealed a massive recent accumulation
 48of TEs equivalent to the size of entire *Drosophila* genomes, which likely explains the large genome
 49sizes in grasshoppers. Although the TE and satDNA repertoires were rather similar between races,
 50the patterns of TE expression and satDNA proliferation suggest rapid evolution of grasshopper
 51genomes on recent timescales.

52Background

53 Eukaryotic genomes exhibit repetitive DNA sequences represented by interspersed
54transposable elements (TEs) and tandem repeats (TRs; e.g., satellite DNA; satDNA), collectively
55known as the ‘repeatome’ [1]. TEs occupy a large fraction of genomes in organisms throughout the
56tree of life [2]. The ubiquity of TEs is driven by the ability to either copy and paste themselves with
57an RNA intermediate step (retrotransposons, class I) or to cut and paste themselves (most DNA
58transposons, class II) within the genome of the host organism [3–5]. Each TE class can be further
59categorized by elements that encode protein products required for transposition (autonomous) and
60those that only contain the sequences (non-autonomous) necessary for trans-recognition by the
61transposition machinery of an autonomous counterpart [6]. Class I elements comprise SINE, LINE
62and LTR. Class II elements comprise TIR (DNA transposons), Crypton, RC (rolling-circle
63replication) and Maverick [7]. The mechanism of transposition allows TEs to invade the genome in
64a parasitic way without general advantage to the individual carrying them [3] but with potentially
65deleterious effects on their host by promoting ectopic recombination, mediating chromosomal
66rearrangements, and disrupting coding sequences [8–10].

67 Another type of repetitive element widely distributed in eukaryotic genomes is satDNA. It
68consists of non-coding repetitive DNA that is tandemly arranged and largely represented in the
69centromeric and pericentromeric heterochromatin of most eukaryotic genomes [11–13]. satDNA
70evolution is influenced by several mechanisms of non-reciprocal genetic exchange such as unequal
71crossing-over, intra-strand homologous recombination, gene conversion, rolling-circle replication
72and transposition [11,14–17]. These mechanisms can gradually increase the copy number of a new
73sequence variant within a satDNA family across the genomes of a sexual population [11,14–
7416,18,19]. Sequences within a satDNA family undergo concerted evolution as repeat exchanges
75occur both within and between members of the satDNA family by non-reciprocal genetic transfers
76between homologous and sometimes non-homologous chromosomes [14,20]. This results in
77frequent homogenization of repeats between copies within species and also between repeat copies

78 located on a same chromosome than between different chromosomes [14,20]. At the same time, the
 79 primary sequence of satDNAs usually mutates quickly, and this rapid satDNA turnover leads to
 80 distinct composition and genomic distribution of satDNAs between strains, populations, subspecies
 81 or species [11,15,18,19,21,22]. The library hypothesis proposes that species do not entirely lose or
 82 gain certain lineages of satDNAs, but, instead, related species share a common collection of
 83 satDNAs that may independently increase or decrease in their copy numbers during or after
 84 speciation [23,24]. Sequence divergence as the outcome of reproductive isolation might then lead to
 85 a formation of species-specific profiles of satDNA sequence variants [25,26]. For grasshoppers and
 86 crickets (Orthoptera), high-throughput sequencing analyses detected 316 satDNA families in ~20
 87 species [12,21,27–32]. Because of the large divergence time between these Orthopteran species (73–
 88 224 millions years, Myr) [33], none of these 316 satDNA families showed interspecific homology,
 89 except for partial sequence homology within satDNAs of *Gryllus* cricket species [27] (unknown
 90 divergence time) and within *Schistocerca* grasshopper species [21] with divergence time < 8 Myr
 91 [34]. Because homology between satDNAs often cannot be detected between distantly related
 92 species due to rapid sequence divergence, an explicit examination of the library hypothesis requires
 93 well-annotated genomes of closely related species.

94 The repeatome is involved in the processes of sex chromosome differentiation [11,35–38].
 95 Both plants and animals have accumulated TEs and satDNAs in the non-recombining regions of Y
 96 and W chromosomes [11,35,37–41]. Studies on neo-Y chromosomes of several *Drosophila* species
 97 proposed that the first steps of Y chromosome degeneration are driven by accumulation of TEs and
 98 satDNAs [36,40]. The non-recombining parts of the Y or W chromosomes may thus expand by
 99 repeat accumulation and heterochromatinization [36,40]. Moreover, transcription of satDNAs
 100 (satRNAs) may have a critical role in centromere function, chromatin silencing, heterochromatin
 101 formation, chromatin modulation, and up-regulation of X-linked genes in chromosome dosage
 102 compensation [42–45]. It has been proposed that interspecific incompatibilities in hybrids between
 103 satRNAs and specific proteins involved in centromere function can contribute to genome

104divergence and the speciation process [25,44,46–48].

105 Grasshoppers generally have large genomes (9 Gb on average, minimum 1.5 Gb and
106maximum 16.6 Gb [49]) likely because of large amounts of repetitive DNA [50,51]. They provide
107ample opportunities to investigate the influence of the repeatome on karyotype evolution because
108grasshoppers are also karyotypically divergent between closely related species [52–54]. However,
109comparative genomic studies in grasshoppers have been hampered by their large genome sizes
110[50,51]. The Australian morabine (Morabinae) grasshopper of the genus *Vandiemenella* (hereafter
111referred to as the *viatica* species group) is a relatively young species complex with estimated
112divergence time $< 0.5\text{--}3.1$ Myr based on a mitochondrial marker [53] and is karyotypically diverse
113[53–55]. It currently contains two nominal species (*V. pichirichi* and *V. viatica*) and five provisional
114species (P24, P25, P45b, P45c, and P50) differentiated by one or more chromosomal
115rearrangements [56,57]. Earlier cytogenetics studies hypothesized that *viatica*19 ($2n = 19$, X0 male)
116is most closely resembling the ancestral karyotype of the *viatica* species group [54,55]. Subsequent
117sequential chromosomal rearrangements, including centric fusions, fissions and inversions, resulted
118in the formation of the present taxa [54,55,58]. Notably, neo-sex chromosomes in the *viatica* species
119group evolved three times independently through recent fusions of the ancestral X chromosome
120with a different autosome each time (P24X0/XY, P25X0/XY and P45bX0/XY races) [54,55]. Given
121that repetitive sequences can play critical roles in the evolution of genome structure and function, a
122comprehensive analysis of TEs and satDNAs of these grasshopper genomes is essential for
123understanding of the genome structure and chromosomal evolution of the *viatica* species group.

124 Here, we characterized TEs and satDNAs in four chromosomal races of the *viatica* species
125group, P24X0, P24XY, P45bX0, and P45bXY, by generating 10X Genomics Chromium linked-read
126data and RNA sequencing data. We then used three complementary methods, namely homology-
127based, structure-based, and de-novo approaches, for annotating the TE and satDNA fraction of the
128genomic reads and assemblies. By comparing sequence divergence within the TE classes, we
129identified the temporal dynamics of TE accumulation. We identified transcriptional activity of many

130subfamilies of TEs in three tissues of these grasshoppers, showing that the TE expression levels
131vary greatly among TEs, tissues and sexes. We also provided evidences that satDNAs expanded and
132contracted in their genomic copy numbers at different time points since the divergence of the
133chromosomal races.

134

135Results

136Genome assembly

137 We sequenced male genomes of four chromosomal races representing two pairs of
138karyotypes with and without neo-sex chromosomes (Figure 1a) using 10X Genomics Chromium
139linked-read libraries with 1,577-1,883 million paired-end reads per library. Average input molecule
140lengths between races ranged from 17.44 to 56.11 kb. By using Supernova 2.1.0 [59], we obtained
141genome assemblies with the following sizes: 3.02 Gb in P24X0, 2.73 Gb in P24XY, 3.27 Gb in
142P45bX0, and 2.94 Gb in P45bXY. The contig N50 for genome assemblies ranged from 29.11 to
14335.69 kb, and the scaffold N50 ranged from 34.85 to 316.69 kb. Chromosomal races with larger
144inferred genome assembly sizes tended to have more fragmented genomes (Table 1). Using BUSCO
145v3 [60] with the Arthropoda dataset, these assemblies had more complete single-copy orthologs (57-
14690 %) than the Illumina assembly of the migratory locust *Locusta migratoria* (Acrididae,
147Oedipodinae) (42 %) [50]. The BUSCO scores also indicated that the fractions of genes that were
148entirely missing in each race (except P45bXY) were lower than what has been observed previously
149in *L. migratoria* genome assembly (Figure 1b). The assembly of P45bXY showed the lowest
150number of complete single copy orthologs across the *V. viatica* genome assemblies but still higher
151than the *L. migratoria* genome assembly [50].

152 Owing to the incompleteness of virtually all animal genome assemblies [61,62], genome
153assembly length tends to be smaller than the actual genome size. To ameliorate this
154underestimation, we inferred genome size from sequencing reads directly by analyzing the
155frequency of k-mers using the findGSE function [63]. The measures of genome sizes varied

between each pipeline used, i.e. between findGSE and the Supernova estimations computed from k-mer distributions (Table 1). We thus report average values between these k-mer-based measures as an approximation of the true genome sizes. This analysis resulted in genome size estimates of 4.26 Gb in P24X0, 4.50 Gb in P24XY, 4.64 Gb in P45bX0, and 5.07 Gb in P45bXY.

160

161 *Transposable element (TE) identification*

162 We first ran RepeatModeler 1.0.8 [64] on each of the four genome assemblies to generate
163 benchmark repeat libraries for annotating the TEs, yielding between 1,361 and 1,398 consensus
164 sequences per race. Between 637 and 668 consensus sequences were initially classified as unknown
165 by RepeatModeler in each library. To further classify these unknown repeats, we manually curated
166 the P24X0 repeat library generated by RepeatModeler because this had the best genome assembly
167 quality based on contig N50 and BUSCO scores. Our manual curation identified 212 new consensus
168 sequences of TE subfamilies in the P24X0 genome (32 % of the unknown repeat consensus
169 sequences above). Next, we used the curated TE consensus sequences of P24X0 to re-classify
170 unknown repeats in the other three libraries by homology searches in RepeatMasker 4.0.8 [65]. This
171 classified a total of 215, 206, and 212 new TE subfamilies (~34 % of the unknown repeat consensus
172 sequences above) in the P24XY, P45bX0, and P45bXY genomes, respectively. From these newly
173 identified TEs, about 2.24-2.69 % of each genome assembly (83-93 subfamilies) were reclassified
174 as LTR retrotransposons and 2.13-2.78 % of each genome assembly (77-98 subfamilies) as DNA
175 transposons. The comparison of TE landscapes, i.e., the distribution of TE-derived bp in bins of
176 Kimura 2-parameter (K2P) distance, in raw RepeatModeler repeat libraries vs. curated libraries
177 highlights the improvement of TE annotation by in-depth manual curation and re-classification
178 (Additional file 2: Figure S1).

179 To further improve the TE annotation, we applied HelitronScanner 1.0 [66] to search for
180 low-copy Helitrons that were missed by RepeatModeler. HelitronScanner and clustering of the
181 nucleotide dataset into clusters that met a similarity threshold of 80 % produced a set of non-

182redundant representative sequences (families) of new Helitrons, i.e. 230, 211, 286, and 221 Helitron
183families in the P24X0, P24XY, P45bX0, and P45bXY genomes, respectively. The bp percentages of
184these new Helitron sequences in the genome assemblies were 10 % (2,743,812 fragments) in
185P24X0, 13 % (2,578,114 fragments) in P24XY, 15 % (3,207,327 fragments) in P45bX0, and 13 %
186(2,837,154 fragments) in P45bXY. We identified a new family of autonomous Helitron
187(“Tukutron”) in the P24X0 genome (Figure 2) in which entire and truncated SINEs were nested in
188its sequence. Tukutron comprises 12,053 fragments in the genome of P24X0 (0.01 % of the
189genome).

190 The above TE annotations revealed that from 66 to 75 % of the assembled genomes of the
191*viatica* species group was composed of TEs (Figure 1c, Table 2). The density of LINEs (715 to 819
192Mb) was highest among the annotated major TE groups, followed by rolling-circle (RC) Helitrons
193(664 to 878 Mb), DNA transposons (679 to 734 Mb), LTRs (160 to 208 Mb) and SINEs (56 to 87
194Mb) (Table 2). Regarding superfamilies of TEs, the most abundant across the assembled genomes
195was RC/Helitron (17-22 %), followed by DNA/TcMar (7%-9%) and LINE/CR1 (7-10 %). Since
196tandem repeats (TRs) were likely underrepresented in genome assemblies, the read-based
197RepeatExplorer2 [67–69] and NOVOplasty 3.7.2 protocols [70] were applied to detect TRs (see
198below). None of the TRs detected using these two approaches were recovered by the above
199RepeatModeler repeat libraries. The relative genomic abundance of detected TRs in the sequenced
200reads was then compared by sampling 4 million read pairs per library and aligning them to the
201combined repeat libraries database (RepeatModeler + HelitronScanner + TRs) with RepeatMasker;
202this quantification was done separately in each race. The analysis revealed that 73 to 79 % of the
203sequenced reads of the *viatica* species group was composed of repeats (Figure 1d, Table 3), and
204showed that, while overall TE proportions were comparable to the assembly proportions, TRs and
205RC/Helitron sequences were better represented in the read-based approach than in the assembly-
206based approach (Figure 1c,d; Table 2 and 3).

207

208 ***Temporal accumulation of TEs***

209 Assuming that the K2P distance from the consensus sequence reflects the time since the
210 insertion of a TE copy, this can be a proxy for the temporal accumulation of TEs [71]. Based on this
211 assumption, we quantified the accumulation of recent TEs that were between 0 and 5 % diverged
212 from the respective consensus sequence (Figure 3). We found a total length of 314-463 Mb of TEs
213 with 0-5 % divergence in each assembly, suggesting massive recent amplifications of the five major
214 TE groups in each race: P24X0 (DNA = 175 Mb, LINE = 126 Mb, SINE = 7 Mb, LTR = 42 Mb,
215 RC = 35 Mb), P24XY (DNA = 167 Mb, LINE = 97 Mb, SINE = 2 Mb, LTR = 25 Mb, RC = 23
216 Mb), P45bX0 (DNA = 200 Mb, LINE = 158 Mb, SINE = 7 Mb, LTR = 56 Mb, RC = 43 Mb) and
217 P45bXY (DNA = 181 Mb, LINE = 122 Mb, SINE = 14 Mb, LTR = 36 Mb, RC = 47 Mb). This is
218 larger than the estimated genome size of *Drosophila* and many other insects [49].

219 We considered the K2P distance bins of 0 to 1 % as very recent TEs that likely accumulated
220 during or after divergence of the chromosomal races. Some of the most abundant TE superfamilies
221 with very recent copies were DNA/DNA, DNA/P, DNA/Sola, DNA/hAT, DNA/TcMar,
222 RC/Helitron, LINE/L2, LTR/LTR, LTR/Gypsy and SINE/tRNA (Figure 3). The numbers of
223 occupied bp of these superfamilies in each assembled genome were different in this divergence bin,
224 with largest variation for LTR/Gypsy (e.g., P24XY = 5 Mb, P24X0 = 9 Mb, P45bXY = 11 Mb,
225 P45bX0 = 21 Mb; coefficient of variation, cv = 170 %) and DNA/TcMar (e.g., P24XY = 3. Mb,
226 P24X0 = 10 Mb, P45bX0 = 22 Mb, P45bXY = 22 Mb; cv = 154 %) indicating differential
227 amplification.

228

229 ***Transcriptional activities of TEs***

230 We next characterized TE expression in the P24X0 and P24XY races (5-11 individuals per
231 tissue/sex/race) by comparing the number of RNA-seq reads mapping to recent TE copies (i.e.,
232 copies with K2P distance < 5 %; hereafter referred to as ‘recent TE expression’ data), using
233 RepEnrich2 (<https://github.com/nerettilab/RepEnrich2>) and DESeq2 1.20.0 [72]. For comparison,

we repeated the analysis with all TE copies regardless of K2P distance. For this analysis, we provided RepEnrich2 with a filtered RepeatMasker annotation file containing only TE loci with < 5% K2P from the consensus sequences. This step was required to survey recent TE expression because RepEnrich2 does not retain locus coordinates, preventing us to subsample recent TE expression if the total TE expression is considered (i.e., including older TE copies with K2P > 5 %).

In the analysis of recent TE expression, there were 598 and 1,415 expressed TE subfamilies with nonzero total read count in reproductive tissues (male testis and female ovary) in P24X0 and P24XY, respectively. Of these, we observed a higher proportion of female-biased TEs (FBTEs) relative to male-biased TEs (MBTEs) in P24X0 (FBTEs 15.05 %, MBTEs 10.37 %); P24XY showed the opposite trend between sexes (MBTEs 4.88 %, FBTEs 3.39 %) (Table 4). We found 574 and 1,415 expressed TE subfamilies with nonzero total read count in head tissues of P24X0 and P24XY, respectively. Of these, we observed a higher proportion of MBTEs relative to FBTEs in P24X0 (MBTEs 3.66 %, FBTEs 0.35 %) and P24XY (MBTEs 1.48 %, FBTEs 0.21 %) (Table 4). In leg tissues, a total of 1,362 and 775 expressed TE subfamilies with nonzero total read count were observed in P24X0 and P24XY, respectively. There was not much variation in TE expression levels in leg between sexes in both races: P24XY (MBTEs 0 %, FBTEs 0.52 %) and P24X0 (MBTEs 0.23 %, FBTEs 0.07 %). The heatmaps showing the expression data of the 50 most highly expressed TE subfamilies in three tissues contain representatives from all five major TE groups (Figure 4). The numbers of differentially expressed TEs between sexes and tissues when including all TE copies regardless of K2P distance were much larger (between 2,354 and 7,240 TEs transcribed; Additional file 1: Table S1) than the recent TE expression data, likely because of transcription of old and inactive TE copies.

256

257 *Tandem repeat (TR) identification and sequence characterization*

Since TRs were underrepresented in the aforementioned linked-read assemblies, we applied two read-based approaches, RepeatExplorer2 [67–69] and NOVOplasty 3.7.2 [70], and merged the

260 outputs of these two to detect TRs in each race. We identified 56 TR families in P24X0, 60 TR
 261 families in P24XY, 71 TR families in P45bX0 and 92 TR families in P45bXY. These TRs included
 262 multigene families (45S rRNA, 5S rRNA, U snRNA, and histones genes), the tandem telomere
 263 repeat (TTAGG)_n, and satDNA families. None of the TRs were found in the aforementioned
 264 RepeatModeler libraries. The total abundance of TRs composed of multigene families, telomere
 265 repeats and satDNAs represented 7.82 %, 6.06 %, 7.53 % and 8.85 % of the sequenced reads of
 266 P24X0, P24XY, P45bX0, and P45bXY, respectively (Table 3, Additional file 1: Table S2-S5).
 267 Compared to these read-based TR quantifications, the abundance of TRs in the genome assemblies
 268 was much lower (i.e. << 0.1 % per assembly), suggesting that most satDNA copies were collapsed
 269 during the assembly process. Read-based approaches thus proved essential to the identification and
 270 quantification of TRs (Table 2 and 3).

271 The telomere repeat was more abundant in the XY races (P24XY and P45bXY, 1.61 % and
 272 1.21 % of the sequenced reads, respectively) than the X0 races (P24X0 and P45bX0, < 0.91 % of
 273 the sequenced reads) likely through independent amplification of the telomere repeats in the two
 274 former (Figure 5). The 45S rRNA gene underwent slightly more differential amplification in P24XY
 275 (abundance 0.61 %) than the other three races (abundances < 0.31 %). Most copies of the telomere
 276 repeat and the 45S rRNA gene were in the divergence bins of K2P < 5 %, suggesting recent
 277 amplification.

278 The satDNAs were the dominant TRs in all four races (Figure 5). RepeatExplorer2 found 45
 279 satDNA families in P24X0 (7 % of sequenced reads), 48 satDNA families in P24XY (4 % of
 280 sequenced reads), 60 satDNA families in P45bX0 (6 % of sequenced reads), and 81 satDNA
 281 families in P45bXY (7 % of sequenced reads) (Additional file 1: Table S2-S5; Additional file 2:
 282 Figure S2). The repeat unit length (monomers) of the satDNA families ranged from 21 to 1,740 bp
 283 in P24X0, from 7 to 1,690 bp in P24XY, from 13 to 1,710 bp in P45bX0, and from 13 to 1,250 bp in
 284 P45bXY. Multiple sequence alignments and dotplots showed that most satDNA families were
 285 present in multiple contigs within the clusters of the RepeatExplorer2 results, with monomers

286differing by at least one nucleotide from each other, indicating that distinct sequence variants of
287monomers and higher-order repeat structures (HORs) were present in each genome (Additional file
2881: Table S2-S5; Additional file 2: Figure S3).

289

290 **Assessment of the satDNA library hypothesis**

291 Because RepeatExplorer2 can only detect satDNAs with abundance > 0.01 % of the genome
292[67–69], homologous satDNAs present below this threshold can remain undetected across samples.
293Thus, we used two approaches to search for homologous satDNAs between races. First, we built a
294satDNA database by concatenating all the satDNA consensus sequences detected in each race
295(totaling 234 consensus sequences), and performed an all-against-all comparison of the consensus
296sequences in the database using the `rm.homology.py` script (<https://github.com/fjruizruano/ngs->
297protocols) [31]. Second, we mapped the genomic reads of each race separately to the satDNA
298database using RepeatMasker. These approaches defined 129 satDNA families across the four races
299(hereafter referred to as satDNA-1 to satDNA-129). Out of the 129 satDNAs, 102 satDNAs were
300shared between all four races, and 27 satDNAs were present or absent in one or more race
301(Additional file 1: Table S6-S7). From these 27 satDNAs families, there were 1-2 satDNA families
302specific to just one race each, the largest number in P24XY and P45bXO (Figure 6a).

303 From the 102 satDNA families that were shared between all four races, there were 50
304satDNA families that showed $cv > 80$ % (cv min 82 % and cv max 199 %) for abundance data
305between races (Additional file 1: Table S8), indicating differential amplification since the
306divergence of the chromosomal races. From these 50 satDNA families, we randomly selected 10
307families to generate a repeat landscape of their relative genomic read abundance in 1 % bins of K2P
308distance. Differential amplification indicated by the differences in abundance and K2P distance
309distribution of sequences was significant for eight satDNAs (Kruskal-Wallis test $P < 0.03$; satDNA-
3101, satDNA-2, satDNA-8, satDNA-17, satDNA-27, satDNA-45, satDNA-47, satDNA-90) but not for
311the remaining two ($P > 0.33$; satDNA-46 and satDNA-111) (Figure 6, Table 5). Among these 10

312satDNAs, the most abundant family was satDNA-1 in P24X0 (average 1.48 %) and the least
313abundant one was satDNA-47 in P24X0 (average K2P << 0.01 %). The most divergent family was
314satDNA-45 in P45bXY (average K2P 25.53 %) and the least divergent one was satDNA-8 in
315P45bXY (average K2P 4.01 %) (Table 5).

316

317Discussion

318Genome assembly

319 The genomes of four chromosomal races of the *viatica* species group that we assembled here
320with 2.94-3.27 Gb assembly sizes are the third largest assembled insect genomes so far, with the
321largest being the two locust grasshoppers *L. migratoria* and *Schistocerca gregaria* with 6.5 and 8.6
322Gb assembly sizes, respectively [50,51]. We believe that genome size estimates based on k-mer
323analysis better represent the genome size of these grasshoppers than the genome assembly sizes
324because highly repetitive sequences (e.g., centromeres, telomeres, satDNAs, non-recombining part
325of sex chromosomes) are likely collapsed during the assembly process [61,62,73,74]. However, the
326two different k-mer approaches yielded quite different estimates between the chromosomal races
327(3.30-3.82 Gb by Supernova, and 5.42-6.32 Gb by findGSE), the reasons for which remain unclear.
328The large differences in contig size (contig N50 29.11-35.69 kb) of the assembled genomes of the
329*viatica* species group, *L. migratoria* (contig N50 9.30 kb) [50] and *S. gregaria* (contig N50 12.03
330kb) [51] are probably due to the difference in the DNA library preparation and sequencing methods.
331The *L. migratoria* genome is based on Illumina short reads (2x 45-150 bp paired-end) while we
332used linked reads (150 bp paired-end Illumina short reads with barcode information from long input
333DNA molecules), and the *S. gregaria* assembly used paired-end and mate-pair Illumina short reads
334and PacBio long reads. The scaffold N50 (158 kb) of the *S. gregaria* genome assembly (8.6 Gb) is
335smaller than the scaffold N50 (317 kb) that we obtained in the P24XO race. In addition, the *S.*
336*gregaria* genome assembly is even more fragmented than the *L. migratoria* as indicated by the
337BUSCO scores (see [51]), indicating that assembling such large grasshopper genomes is

challenging even using the combination of technologies above. Although the smaller genome assemblies of the *viatica* species group were more contiguous than the assembly of *L. migratoria*, they still contained missing and fragmented genes. This may be due to 1) large numbers of repetitive non-coding DNAs, 2) intron gigantism, 3) errors during the assembly process, 4) true missing genes, 5) failure in identifying any significant matches, and/or 6) failure in the gene prediction step to produce even a partial gene model that might have been recognized as a fragmented BUSCO match [60].

TE dynamics and genome evolution

Approximately 66 to 72 % of the genome assembly of the *viatica* species group corresponds to TEs. These are even larger than the 60 to 62 % reported for the genome assemblies of *L. migratoria* [50] and *S. gregaria* [51] respectively, likely owing to the combination of methodological approaches to annotate TEs that we used. Given the deep divergence of Eumastacoidea and Acridoidea (~197 mya [33]), the two Orthopteran superfamilies to which the *viatica* species group and the locusts grasshoppers belong, respectively, these findings suggest that large repeatomes are widespread in grasshoppers. The *S. gregaria* genome assembly has 18,815 annotated genes, and the *L. migratoria* genome assembly has a similar number of annotated genes (17,307) to those recently reported in two cricket species with genome assembly sizes of 1.6 Gb (TE content 40 %; [75]), indicating the absence of partial or whole genome duplication events in these orthopteran lineages. Additionally, there was no evidence for such large-scale genome duplication events in the *viatica* species group because only 3-5 % of BUSCO genes were duplicated. Therefore, the large genome sizes in these grasshoppers are likely due to the expansion of TEs, which has been correlated with genome size evolution across the Tree of Life [2,76]. Indeed, we found massive recent amplification in hundreds of Mb (between 314 and 463 Mb) of the TE groups per genome assembly, an amount that is notably larger than the estimated genome size of many other insects [49,76]. The recent amplification mainly occurred in eight TE superfamilies

(DNA/DNA, DNA/P, DNA/Sola, DNA/hAT, DNA/TcMar, RC/Helitron, LINE/L2, LTR/LTR, LTR/Gypsy and SINE/tRNA) with largest variation for LTR/Gypsy and DNA/TcMar, indicating that the recent amplifications of TE superfamilies have widely shaped the TE landscape in the genomes of the chromosomal races. We thus suggest that the massive proliferation of TEs combined with a slow deletion rate might contribute to the genomic gigantism in grasshoppers, as proposed for other large eukaryotic genomes [77].

370

371 **Recently active TEs are permissively transcribed in gonads**

372 We restricted our analysis to transcripts that originated from recent TEs (i.e., $K2P < 5\%$) because these are likely to be relevant source of transcriptional and transpositional activity. Our results demonstrated that recently active TE superfamilies from all five major TE groups (LINE/SINE/DNA/RC/LTR) are transcribed, suggesting that at least some of the transcribed TEs are capable of (retro)transposition. Recent TE expression tended to be differentially expressed in gonads compared to somatic tissues. This might indicate that TEs are transcribed and transpose themselves more frequently in gonads, transmitting new TE insertions to the next generation. The grasshopper ovaries and testes showed uneven expression of recent TEs, suggesting that substantial TE transcriptional variation likely exists across sexes. The expression variation of recent TEs between reproductive and somatic tissues is puzzling. We speculate that this might result from either global epigenetic reprogramming during gametogenesis or the many more cell types/stages present in gonads than in individual somatic tissues of these grasshoppers. Alternatively, TE control might be tighter in somatic tissues, such that tight repression of TEs is important for the host and more feasible in somatic tissues (no global epigenetic reprogramming). It remains to be investigated if higher transcriptional activities of all five major TE groups, particularly in reproductive tissues, is associated with higher TE repressive mechanism activation (piwi/piRNA pathway [78–80]) to prevent the potentially deleterious effects of TE (retro)transposition in the host during global epigenetic reprogramming [78–81].

390

391The chromosomal races of *V. viatica* species group share a common collection of satDNAs
392which mostly experienced quantitative changes during evolution

393 To our knowledge, we uncovered the largest collection of satDNA families (129) ever
 394reported for eukaryotic genomes. We propose that the 102 satDNA families shared among all four
 395chromosomal races (Additional file 1: Table S6 and S8) represent the “library” present in the
 396*viatica* ancestor. The 27 satDNA families that were not shared between all the four races either
 397emerged after the divergence of the *viatica* ancestor or were lost in one or more races. This implies
 398that the essential step in the evolution of a satDNA family might either be the acquisition of
 399biological functions or the accumulation of sufficiently many copies to be maintained in the
 400“library” over long evolutionary periods. How the novel satDNA families emerged remains unclear
 401in the *viatica* species group, although unequal crossing-over, intra-strand homologous
 402recombination, gene conversion, rolling-circle replication, and transposition are possible
 403mechanisms [11,14–16,18,20,22]. After satDNA emergence, it is logical to assume that satDNA
 404families stochastically expand or disappear and are only maintained in the long term if they acquire
 405a function, such as in centromeres or heterochromatin formation. To test for satDNA functionality
 406and to determine whether satDNAs were independently acquired or lost, additional data (e.g.
 407chromosome *in situ* hybridization and ChIP-seq) is needed including races/species from the other
 408morabine grasshoppers and under a robust phylogenetic hypothesis.

409 In line with the satDNA library hypothesis [23,24], 50 of the 102 satDNA families shared
 410among all four chromosomal races experienced quantitative changes in copy number, and these
 411happened over different K2P bins of divergence (see Figure 6b-d), suggesting rapid satDNA
 412proliferation. The changes in copy number likely occurred by unequal crossing over, which is the
 413mechanism that can yield changes in TR abundance, either as gains (amplifications) or losses
 414(contractions) [82]. The large number of satDNA families in these chromosomal races is puzzling.
 415The non-coding satDNAs have been traditionally viewed as mostly useless material capable of

416accumulating primarily in heterochromatin [11–13,27,29] until they become a too heavy load for
417the host genome (reviewed in [45]). It will be interesting to test whether the differential
418amplification of satDNAs is correlated with the amount of heterochromatin. On the other hand, it
419has been suggested that differential expression of satDNAs as satRNAs can cause genomic
420incompatibilities in hybrids because satRNAs play critical roles in kinetochore assembly (i.e. by
421binding to specific centromeric proteins like CENP-A and CENP-C) [25,26,45,46], heterochromatin
422formation [44,47,48] and function during cell division via siRNAs and piRNA pathways in
423*Schizosaccharomyces pombe*, *Drosophila*, nematodes, humans [44–48,83]. The satDNA families
424identified here are thus a set of candidates for future studies on which satDNAs are located in
425centromeres and which are involved in heterochromatin formation of grasshoppers.

426

427Conclusion

428 We have generated the so far most contiguous genome assemblies of grasshoppers with 66-
42975 % repeat contents using 10X Genomics linked-read sequencing. In-depth repeat annotation
430proved essential to elucidate the composition and characteristics of TEs and satDNAs in large and
431repetitive genomes such as grasshoppers. We showed a massive recent proliferation of a wide range
432of TEs, many of which are transcribed more frequently in germline than somatic tissues. In
433addition, we uncovered the largest number of satDNA families ever reported in eukaryotic
434genomes, and showed that, despite the recent divergence of the four *viatica* chromosomal races,
435satDNA evolution underwent rapid expansions or contractions in copy number.

436

437Methods

438Taxon sampling, DNA and RNA extraction, and sequencing

439 Chromosomal races of *Vandiemena* morabine grasshoppers were collected between 2002
440and 2017 in South Australia. To identify the races, the testes were dissected from males and fixed
441for karyotyping as described previously [53,84]. The remaining body parts were flash-frozen in

liquid nitrogen and stored at -80 °C in the Australian Biological Tissue Collection until subsequent DNA extraction. One male per race (P24X0, P24XY, P45bX0 and P45bXY) was used for DNA extraction from either heads or legs using the MagAttract HMW DNA Kit (QIAGEN, Hilden, Germany; Cat No. 67563). Sequencing libraries were prepared as recommended by Chromium Genome preparation kit (10X Genomics, Inc., Pleasanton, CA, USA; Cat No. 120215). Paired-end read sequencing (2x150 bp) was performed on the Illumina HiSeq X (Illumina, Inc., San Diego, CA, USA) using the Chromium library.

Tissues from males (head, leg, and testes) and females (head, leg, and ovary) of the chromosomal races P24X0 and P24XY were dissected and placed in RNAlater (ThermoFisher Scientific, Waltham, MA, USA) and stored at -80 °C until subsequent RNA extraction. In total, 42 males and 62 females were used for RNA extraction (5-11 individuals per tissue/sex/race). We extracted RNA with phenol-based phase separation using the TRIzol reagent (ThermoFisher Scientific) following the standard protocol recommended by the supplier. Sequencing libraries were prepared according to the TruSeq stranded mRNA library preparation kit (Illumina, Inc., Cat No. 20020594/5) including poly-A selection. Paired-reads (2x100 bp) were sequenced on the NovaSeq 6000 S2 flowcell (Illumina, Inc.). DNA and RNA library preparation and sequencing were performed at the SNP&SEQ Technology Platform in Uppsala, Uppsala University, Department of Medical Sciences, Uppsala Biomedical Centre (BMC) (Uppsala, Sweden).

Genome size estimation and genome assembly

We performed genome size estimation by counting k-mer frequency of the quality checked 10X Genomics linked-reads. The multiple fastq.gz files were two-step processed for counting k-mers using Jellyfish 2.2.6 [85] with the following setting: -t 8 -C -m 18 -s 5G --min-quality=20 --quality-start=33. We used the R package findGSE [63] to estimate genome sizes by using the output of Jellyfish.

We used Supernova 2.1.0 [59] to generate a “pseudohaploid” genome assembly of males of

468P24X0, P24XY, P45bX0 and P45bXY using 10X Genomics linked-read data. Supernova uses the
46910X Genomics linked reads generated from a single library of DNA from an individual organism as
470source, potentially allowing for the assembly of longer contigs and scaffolds than conventional
471short-read technologies [59]. For comparative analysis of the generated draft assemblies, we
472downloaded the *Locusta migratoria* genome from the NCBI database (GCA_000516895.1) [50].
473The completeness of the assemblies was evaluated with BUSCO v3 [60] with the Arthropoda
474database [86] as a reference. The recovered matches were classified as complete if their lengths
475were within the expectation of the BUSCO profile match lengths. If these were found more than
476once, they were classified as duplicated. The matches that were only partially recovered were
477classified as fragmented, and BUSCO groups that passed the test of gene prediction but for which
478there were no matches in the database were classified as missing.

479

480Assessing the TEs content across genomes

481 We used the draft male genome assemblies of P24X0, P24XY, P45bX0 and P45bXY to
482generate repeat libraries for each of the four genomes using RepeatModeler 1.0.8 [64]. Because the
483P24X0 had the highest assembly contiguity and BUSCO scores, the repeats classified as unknown
484in the P24X0 RepeatModeler library were selected for manual curation following the method used
485in Suh et al. [87]. Every consensus sequence was aligned back to the assembled genome sequence
486of P24X0, then the best 20 BLASTn hits were collected, extended by 2 kb and aligned to one
487another using MAFFT 7 [88]. Manually curated consensus sequences of P24X0 were then classified
488as TE families/subfamilies based on the proposed classification system for TEs (open reading
489frames, terminal repeats, target site duplications) [7] and Repbase similarity searches [89]. We
490removed redundancies from the curated TE consensus sequences of P24X0 by merging sequences
491that were greater than 80% similar using CD-HIT-EST [90], with the following setting: -c 0.80 -n 5
492-M 0 -aS 0.80 -G 0 -g 1.

493 The curated TE library from P24X0 race was used as a reference to re-classify repeats in the

TE libraries of the other three races (i.e., P24XY, P45bX0 and P45bXY) by using RepeatMasker 4.9.0 [65]. We removed redundancies of the re-classified TEs within each library using CD-HIT-EST. We thus obtained the following combinations of libraries: 1) P24X0 RepeatModeler library (RML) + Arthropoda Repbase library (ARL) + the curated library, 2) P24XY RML + ARL + curated library, 3) P45bX0 RML + ARL + curated library, and 4) P45XY RML + ARL + curated library. We used these libraries to mask their respective genomes to annotate repeats using RepeatMasker [65]. We then used each of the .align RepeatMasker output files per genome to estimate abundance and divergence of each TE superfamily. To visualize the temporal activity/accumulation of TEs across races, we generated landscape bar plots depicting the relative abundance of repeat elements on the y-axis and the K2P distance from the respective consensus sequence on the x axis.

In addition to the annotation of TEs in the four repeat libraries above, we also searched for new Helitrons across the genome assemblies using HelitronScanner 1.0 [66]. HelitronScanner uses the two-layered local combinatorial variable (LCV) tool for Helitron identification. To avoid false positive in Helitron searches, we used thresholds of 7 for both Helitron ends (14 for sum of LCV scores) as a parameter in the HelitronScanner runs. We used CD-HIT-EST to remove redundancies of the detected Helitron sequences by merging sequences that were greater than 80 % similar. This produced a set of non-redundant representative sequences (families) of new Helitrons. We used the detected Helitrons in each chromosomal race in combination with the libraries above mentioned to finally mask the respective genomes.

We further quantified the repeatome between races using read-based approaches to compare with the assembly-based approach. For this purpose, the relative genomic abundance of repeats in the sequenced reads were compared by random sampling of 4 million read pairs per library and masking them with the combined repeat libraries (RepeatModeler + HelitronScanner + TRs) using RepeatMasker. This quantification was done separately in each race using their respective repeat library. TRs were detected using RepeatExplorer2 [67–69] and NOVOplasty 3.7.2 [70] (see below).

520 **Transcription of TEs**

521 We used Illumina RNA-seq reads (2x 100 bp) from males and females (testis, ovary, head
522 and leg) of P24X0 and P24XY races to investigate the transcriptional profile of TEs in each tissue.
523 For comparative purposes, we used the RepeatMasker annotation (with simple and low complexity
524 repeats removed) from the P24X0 and P24XY races to check whether the TEs show differential
525 expression among sexes and tissues, and to compare the intensity and direction of the bias among
526 races. RNA-seq reads per sample were trimmed with TrimGalore ([https://github.com/FelixKrueger/](https://github.com/FelixKrueger/TrimGalore)
527 TrimGalore). Erroneous k-mers from Illumina paired-end reads were removed using rCorrector
528 [91]. Unfixable reads (often riddled with N nucleotides or represented by other low complexity
529 sequences) were discarded using FilterUncorrectablePEfsta.py script obtained from the Harvard
530 Informatics [GitHub](https://github.com/harvardinformatics/TranscriptomeAssemblyTools) repository
531 (<https://github.com/harvardinformatics/TranscriptomeAssemblyTools>). We used SortMeRNA 3.0.3
532 [92] for local alignment, filtering, mapping, and clustering to remove rRNA.

533 We applied the RepEnrich2 protocol (<https://github.com/nerettlab/RepEnrich2>) to estimate
534 TE transcript expression levels in each genome assembly. To assign mapped reads to a genomic
535 locus, RepEnrich2 requires an annotation file that specifies repeat element coordinates. The
536 RepeatMasker annotation was used for the approach. We estimated the expression levels of reads
537 that originated from recent TE copies. For this analysis, we provided RepEnrich2 with a filtered
538 RepeatMasker annotation containing only TE copies with < 5 % K2P from the consensus
539 sequences. The filtered RepeatMasker annotation was then used to build the final annotation
540 required for RepEnrich2 protocol by running the RepEnrich2_setup script. Next, the pre-processed
541 RNA-seq reads were mapped to each of the assemblies using Bowtie2 2.2.9 [93] as the read
542 mapper. The quantification step included raw read count and differential expression analysis of TEs
543 estimated by the RepEnrich2 protocol and DESeq2 1.20.0 [72], respectively. The data regarding
544 sex-biased TE expression were compared between gonads, heads, and legs from males and females
545 of the chromosomal races P24X0 and P24XY. The statistical analysis of differentially expressed TE

was performed using DESeq2 as implemented in the Bioconductor package [94] in R [95]. All P -values were adjusted using the Wald test as implemented in DESeq2. A TE was considered biased if the comparison for the factor condition (samples) yielded an adjusted $P < 0.05$. The degree of bias was determined by the log2 fold-change (log2FC) difference between conditions as calculated in DESeq2, i.e., those TEs with $\log_2\text{FC} > 0$ and $\log_2\text{FC} < 0$ and with an adjusted $P < 0.05$ were considered as male-biased TE (MBTEs) and female-biased TE (FBTEs), respectively.

Tandem repeat detection in sequencing reads

We used RepeatExplorer2 [67–69] to identify satDNAs using the linked reads of each chromosomal race used for genome assembly above. RepeatExplorer was run separately in each race. Prior to RepeatExplorer2 graph-based clustering analysis, sequence reads were quality-trimmed using TrimGalore. The trimmed paired-end reads were joined by using the “fastq-join” software of the FASTX-Toolkit suit [96] with default parameters. The joined paired-end reads were then subject to graph-based clustering and assembly using RepeatExplorer2. A set of randomly selected 2,000,000 Chromium linked reads with average length of 150 bp were used as an input for clustering analysis.

We used the dotplot graphic alignment tool implemented in Dotlet [97] to confirm the tandem organization of those clusters with high graph density in RepeatExplorer2 output. Additionally, we used the FlexiDot 1.06 [98] suite to generate all-against-all dotplots visualization of those clusters identified as tandem arrays. The monomer with maximum length was used as the representative copy for each satDNA family, and as the query sequences in further BLASTn (<http://www.ncbi.nlm.gov/Blast/>) and Repbase [89] searches to check for similarity with published sequences, and as a query to check for overlap with the RepeatModeler libraries. We named each satDNA family as Vv (from *Vandiemenna viatica* group) following by the provisional taxon name (i.e., P24X0, P24XY, P45bX0 and P45bXY) and a number in descending order of the genome proportion. To search for HORs, we counted the maximum number of tandem monomer arrays per

contig for each satDNA family by analyzing the dotplots. We then counted the total number of monomers present in each cluster. The multiple sequence alignments of satDNA copies were generated using Muscle [96] implemented in MEGA 5 [99].

We ran NOVOPlasty 3.7.2 [70] to search for tandem multigene families, such as 45S rRNA, 5S rRNA, U snRNA (U1, U2, U5, U6), and histone genes (H1, H2A, H2B, H3 and H4). We used known seed sequences of these genes coming from other species and k-mer between 21 and 23 as parameters for assembly: *Gomphocerinae* sp. for the 45S rRNA gene (AY859546.1), *Ronderosia bergi* for the 5S rRNA gene (KP213274), *Nasonia vitripennis* for the H1 histone gene (XM_003423983.3), *Culex quinquefasciatus* for the H2A (XM_001862631.1) and H2B (XM_001870471.1) histone genes, *Locusta migratoria* for the H3 histone gene (GU111931.1), *N. vitripennis* for the H4 histone gene (XM_001599180.3:49-357), *Eyprepocnemis plorans* for the U1 snRNA gene (KJ606069.1), *Abracris flavolineata* for the U2 rRNA gene (KP975085.1), and *D. melanogaster* for the U5 snRNA (NR_001933.1) and U6 snRNA genes (NR_002081.1). The canonical tandem telomere repeat (TTAGG)_n recognized in insects [100] was recovered from the RepeatExplorer2 output.

The relative genomic abundances and K2P divergences of TRs were estimated individually in each race by sampling 4 million read pairs per library retrieved by seqtk (<https://github.com/lh3/seqtk>) and aligning them to the TR database with RepeatMasker. The sampled reads were mapped to dimers of satDNA consensus sequences, and for smaller TRs, several monomers were concatenated until reaching roughly 150 bp array length. To visualize the temporal accumulation of TRs across races, we generated TR landscape plots that depict the relative abundance of repeat elements on the y axis and the K2P distance from the respective consensus sequence on the x axis.

To search for homologous satDNAs between races, we used two approaches. First, we built a satDNA database by concatenating all the satDNA consensus sequences detected in each race and performed an all-against-all comparison of the consensus sequences in the database using the

598rm.homology.py script (<https://github.com/fjruizruano/ngs-protocols>) [31]. Second, we mapped the
599genomic reads of each race separately to the satDNA consensus database using RepeatMasker, as
600mentioned above, and recorded presence/absence of satDNAs in each of the four races.

601

602Abbreviations

603A+T%: Percentage of A + T bases, i.e., the molar ratio of adenine and thymine in DNA

604cv: Coefficient of variation

605HOR: Higher-order repeat structure

606K2P: Kimura 2-parameter distance

607LINE: Long interspersed element

608LTR: Long terminal repeat

609SINE: Short interspersed element

610TE: Transposable element

611TIR: Terminal inverted repeat

612TR: Tandem repeat

613RC: Rolling-circle transposable element

614satDNA: satellite DNA

615MBTE: Male-biased transposable element

616FBTE: Female-biased transposable element

617log2FC: log2 fold-change

618

619Availability of data and materials

620 The datasets generated as part of the study are available as supplementary information.

621Repeat consensus sequences are available as Additional file 3 (P24X0), Additional file 4 (P24XY),

622Additional file 5 (P45bX0) and Additional file 6 (P45bXY). Each of these additional files contain

623libraries generated by RepeatModeler (including curated and re-classified repeats), HelitronScanner

624and RepeatExplorer2. Raw reads and assemblies were deposited in Sequence Read Archive
625(accession numbers XXXXXXXXX-XXXXXXX)..

626

627Competing interest

628 The authors declare that they have no conflict of interest.

629

630Funding

631 This work was supported by the Swedish Research Council Vetenskapsrådet (grant number
6322014-6325 to TK), Marie Skłodowska Curie Actions, Co-fund Project INCA (grant number 600398
633to TK), and the Swedish Research Council Formas (grant number 2017-01597 to AS). OMPG was
634supported by a postdoctoral fellowship from Sven och Lilly Lawskis fund. The funders had no role
635in the design of the study and collection, analysis, and interpretation of data and in writing the
636manuscript.

637

638Author's contribution

639 Conceptualization: OMPG, TK, AS; Formal analysis: OMPG, JK; Funding acquisition:
640OMPG, TK, AS; Investigation: OMPG, JK, MPF; Project Administration: TK, AS; Resources: TK,
641AS, TMB, SJBC, KKJ; Supervision: TK, AS; Validation: OMPG, JK; Visualization: OMPG;
642Writing – original draft: OMPG; Writing – remaining drafts: OMPG, TK, AS, TMB, SJBC, KKJ.

643

644Acknowledgments

645 The manuscript was improved by comments from the Julie Blommaert, Valentina Peona,
646Francisco J. Ruiz-Ruano and Diogo C. Cabral-de-Mello. TK is grateful for Roger Butlin and the
647University of Sheffield for supporting this research during the INCA fellowship. Computations
648were performed on resources provided by the Swedish National Infrastructure for Computing
649(SNIC) through the Uppsala Multidisciplinary Center for Advanced Computational Science

650(UPPMAX). Sequencing was performed by the SNP&SEQ Technology Platform in Uppsala. The
651facility is part of the National Genomics Infrastructure (NGI) Sweden and Science for Life
652Laboratory. The SNP&SEQ Platform is also supported by the Swedish Research Council and the
653Knut and Alice Wallenberg Foundation.

654

655References

6561. Kim YB, Oh JH, McIver LJ, Rashkovetsky E, Michalak K, Garner HR, et al. Divergence of
657*Drosophila melanogaster* repeatomes in response to a sharp microclimate contrast in Evolution
658Canyon, Israel. *Proc Natl Acad of Sci USA*. 2014;111:10630–5.

6592. Hua-Van A, Le Rouzic A, Boutin TS, Filée J, Capy P. The struggle for life of the genome's selfish
660architects. *Biology Direct*. 2011;6:1–29.

6613. Orgel LE, Crick FHC. Selfish DNA: The ultimate parasite. *Nature*. 1980;284:604–7.

6624. Kidwell MG. Chapter 3 - Transposable Elements. In: Gregory TR, editor. Burlington: Academic
663Press; 2005. p. 165–221.

6645. Kazazian HH. Mobile elements: drivers of genome evolution. *Science*. 2004;303:1626 LP –
6651632.

6666. Suh A. Genome size evolution: small transposons with large consequences. *Curr Biol*.
6672019;29:R241–3.

6687. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification
669system for eukaryotic transposable elements. *Nat Rev Genet*. 2007;8:973–82.

6708. Bennetzen JL. Transposable element contributions to plant gene and genome evolution. *Plant*
671*Mol Biol*. 2000;42:251–69.

6729. Finnegan DJ. Transposable elements. *Current Opinion in Genetics & Development*. 1992;2:861–
6737.

67410. Montgomery EA, Huang SM, Langley CH, Judd BH. Chromosome rearrangement by ectopic
675recombination in *Drosophila melanogaster*: Genome structure and evolution. Genetics.
6761991;129:1085–98.
67711. Charlesworth B, Sniegowski P, Stephan W. The evolutionary dynamics of repetitive DNA in
678eukaryotes. Nature. 1994;371:215–20.
67912. Milani D, Bardella VB, Ferretti ABSM, Palacios-Gimenez OM, Melo A de S, Moura RC, et al.
680Satellite DNAs unveil clues about the ancestry and composition of B chromosomes in three
681grasshopper species. Genes. 2018;9.
68213. Khost DE, Eickbush DG, Larracuente AM. Single-molecule sequencing resolves the detailed
683structure of complex satellite DNA loci in *Drosophila melanogaster*. Genome Res. 2017;27:709–
68421.
68514. Đurđica Ugarković, Miroslav Plohl M. Variation in satellite DNA profiles - Causes and effects.
686EMBO Journal. 2002;21:5955–9.
68715. Dover G. Molecular drive: A cohesive mode of species evolution. Nature. 1982;299:111–7.
68816. Plohl M, Luchetti A, Meštrović N, Mantovani B. Satellite DNAs between selfishness and
689functionality: Structure, genomics and evolution of tandem repeats in centromeric
690(hetero)chromatin. Gene. 2008;409:72–82.
69117. Lopez-Flores I, Garrido-Ramos MA. The repetitive DNA content of eukaryotic genomes.
692Genome Dyn. 2012;7:1–28.
69318. Dover G. Molecular drive. Trends Genet. 2002;18:587–9.
69419. Walsh JB. Persistence of tandem arrays: implications for satellite and simple-sequence DNAs.
695Genetics. 1987;115:553–67.
69620. Garrido-Ramos MA. Satellite DNA: an evolving topic. Genes. 2017;8.
69721. Palacios-Gimenez OM, Milani D, Song H, Marti DA, López-León MD, Ruiz-Ruano FJ, et al.

- 698eight million years of satellite DNA evolution in grasshoppers of the genus *Schistocerca* illuminate
699the ins and outs of the library hypothesis. *Genome Biol Evol.* 2020;12:88–102.
70022. Smith GP. Evolution of repeated DNA sequences by unequal crossover. *Science.* 1976;191:528–
70135.
70223. Salser W, Bowen S, Browne D, el-Adli F, Fedoroff N, Fry K, et al. Investigation of the
703organization of mammalian chromosomes at the DNA sequence level. *Fed Proc.* 1976;35:23–35.
70424. Fry K, Salser W. Nucleotide sequences of HS-alpha satellite DNA from kangaroo rat
705*Dipodomys ordii* and characterization of similar sequences in other rodents. *Cell.* 1977;12:1069–84.
70625. Ferree PM, Prasad S. How can satellite DNA divergence cause reproductive isolation? Let us
707count the chromosomal ways. *Genetics Res Int.* 2012;2012:430136–430136.
70826. Gallach M. Recurrent turnover of chromosome-specific satellites in *Drosophila*. *Genome Biol*
709*Evol.* 2014;6:1279–86.
71027. Palacios-Gimenez OM, Bardella VB, Lemos B, Cabral-De-Mello DC. Satellite DNAs are
711conserved and differentially transcribed among *Gryllus* cricket species. *DNA Res.* 2018;25:137–47.
71228. Palacios-Gimenez OM, Milani D, Lemos B, Castillo ER, Martí DA, Ramos E, et al. Uncovering
713the evolutionary history of neo-XY sex chromosomes in the grasshopper *Ronderosia bergii*
714(Orthoptera, Melanoplinae) through satellite DNA analysis. *BMC Evol Biol.* 2018;18:2–2.
71529. Ferretti ABSM, Milani D, Palacios-Gimenez OM, Ruiz-Ruano FJ, Cabral-de-Mello DC. High
716dynamism for neo-sex chromosomes: satellite DNAs reveal complex evolution in a grasshopper.
717*Heredity.* 2020;125:124–137.
71830. Ruiz-Ruano FJ, Castillo-Martínez J, Cabrero J, Gómez R, Camacho JPM, López-León MD.
719High-throughput analysis of satellite DNA in the grasshopper *Pyrgomorpha conica* reveals
720abundance of homologous and heterologous higher-order repeats. *Chromosoma.* 2018;127:323–40.
72131. Ruiz-Ruano FJ, López-León MD, Cabrero J, Camacho JPM. High-throughput analysis of the

- 722satellitome illuminates satellite DNA evolution. *Sci Rep.* 2016;6:28333–28333.
72332. Ruiz-Ruano FJ, Cabrero J, López-León MD, Camacho JPM. Satellite DNA content
724illuminates the ancestry of a supernumerary (B) chromosome. *Chromosoma.* 2016;1–14.
72533. Song H, Amendt C, Cigliano MM, Desutter-Grandcolas L, Heads SW, Huang Y, et al.
726300 million years of diversification: elucidating the patterns of orthopteran evolution based on
727comprehensive taxon and gene sampling. *Cladistics.* 2015;31:621–51.
72834. Song H, Foquet B, Mariño-Pérez R, Woller DA. Phylogeny of locusts and grasshoppers reveals
729complex evolution of density-dependent phenotypic plasticity. *Sci Rep.* 2017;7:1–13.
73035. Matsunaga S. Junk DNA promotes sex chromosome evolution. *Heredity.* 2009;102:525–6.
73136. Steinemann S, Steinemann M. Retroelements: Tools for sex chromosome evolution. *Cytogenet*
732*Genome Res.* 2005;110:134–43.
73337. Matsubara K, Tarui H, Toriba M, Yamada K, Nishida-Umehara C, Agata K, et al. Evidence for
734different origin of sex chromosomes in snakes, birds, and mammals and step-wise differentiation of
735snake sex chromosomes. *Proc Natl Acad Sci USA.* 2006;103:18190–5.
73638. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of
737the chicken genome provide unique perspectives on vertebrate evolution. *Nature.* 2004;432:695–
738716.
73939. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, et al. The
740male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature.*
7412003;423:825–37.
74240. Bachtrog D, Mahajan S, Bracewell R. Massive gene amplification on a recently formed
743*Drosophila* Y chromosome. *Nature Ecol Evol.* 2019;3:1587–97.
74441. Zhang H, Kobližková A, Wang K, Gong Z, Oliveira L, Torres GA, et al. Boom-bust turnovers of
745megabase-sized centromeric DNA in solanum species: rapid evolution of DNA sequences

- 746associated with centromeres. *The Plant Cell*. 2014;26:1436–47.
74742. Joshi SS, Meller VH. Satellite repeats identify X chromatin for dosage compensation in
748*Drosophila melanogaster* males. *Curr Biol* 2017;27:1393-1402.e2.
74943. Gallach M. 1.688 g/cm³ satellite-related repeats: a missing link to dosage compensation and
750speciation. *Mol Ecol*. 2015;24:4340–7.
75144. Ferree PM, Barbash DA. Species-specific heterochromatin prevents mitotic chromosome
752segregation to cause hybrid lethality in *Drosophila*. *PLoS Biol*. 2009;7:e1000234–e1000234.
75345. Kuhn GCS. Satellite DNA transcripts have diverse biological roles in *Drosophila*. *Heredity*.
7542015;1–2.
75546. Henikoff S, Ahmad K, Malik HS. The centromere paradox: stable inheritance with rapidly
756evolving DNA. *Science*. 2001;293:1098–102.
75747. Barbash DA. Genetic testing of the hypothesis that hybrid male lethality results from a failure in
758dosage compensation. *Genetics*. 2010;184:313–6.
75948. Brown JD, O'Neill RJ. Chromosomes, conflict, and epigenetics: chromosomal speciation
760revisited. *Ann Rev Genom Human Genet*. 2010;11:291–316.
76149. Gregory TR. Animal Genome Size Database. Available from: <http://www.genomesize.com>
76250. Wang X, Fang X, Yang P, Jiang X, Jiang F, Zhao D, et al. The locust genome provides insight
763into swarm formation and long-distance flight. *Nat Commun*. 2014;5:2957–2957.
76451. Verlinden H, Sterck L, Li J, Li Z, Yssel A, Gansemans Y, et al. First draft genome assembly of
765the desert locust, *Schistocerca gregaria*. *F1000Res*. 2020;9:775.
76652. Castillo ER, Marti DA, Bidau CJ. Sex and neo-Sex chromosomes in Orthoptera: a Review. *J*
767*Orthoptera Res*. 2010;19:213–31.
76853. Kawakami T, Butlin RK, Adams M, Saint KM, Paull DJ, Cooper SJB. Re-examination of a

- 769proposed case of stasipatric speciation: phylogeography of the Australian morabine grasshoppers
770(*Vandiemenella viatica* species group). Mol Ecol. 2009;18:3429–42.
77154. Hewitt GM. Animal Cytogenetics 3: Insecta 1, Orthoptera. Berlin: Gebruder Borntraeger:
772Berlin; 1979. p. 170.
77355. White MJD. Animal cytology and evolution. Cambridge, England: University Press; 1973.
77456. White MJD, Blackith RE, Blackith RM Cheney J. Cytogenetics of the *viatica* group morabine
775grasshoppers. I. The coastal species. Aust J Zool. 1967;15:263–302.
77657. MJD White, KHL Key, M Andre and J Cheney. Cytogenetics of the *viatica* group of morabine
777grasshoppers II. Kangaroo Island populations. Aust J Zool. 1969;17:313–28.
77858. White MJD. Modes of Speciation. Freeman W, editor. San Francisco, CA, USA;
77959. Marks P, Garcia S, Barrio AM, Belhocine K, Bernate J, Bharadwaj R, et al. Resolving the full
780spectrum of human genome variation using Linked-Reads. Genome Res. 2019;29:635–45.
78160. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing
782genome assembly and annotation completeness with single-copy orthologs. Bioinformatics.
7832015;31:3210–2.
78461. Peona V, Weissensteiner MH, Suh A. How complete are “complete” genome assemblies?—an
785avian perspective. Mol Ecol Res. 2018;18:1188–95.
78662. Peona V, Blom MPK, Xu L, Burri R, Sullivan S, Bunikis I, et al. Identifying the causes and
787consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise.
788bioRxiv. 2019;2019.12.19.882399-2019.12.19.882399.
78963. Sun H, Ding J, Piednoël M, Schneeberger K. findGSE: estimating genome size variation within
790human and *Arabidopsis* using k-mer frequencies. Bioinformatics. 2018;34:550–7.
79164. Smit AFA, Hubley R. RepeatModeler. Available from:
792<http://www.repeatmasker.org/RepeatModeler/>

79365. Smit AFA, Hubley R, Green P. RepeatMasker. Available from: <http://www.repeatmasker.org>
79466. Xiong W, He L, Lai J, Dooner HK, Du C. HelitronScanner uncovers a large overlooked cache of
795Helitron transposons in many plant genomes. *Proc Natl Acad Sci USA*. 2014;111:10263–8.
79667. Novák P, Neumann P, Macas J. Graph-based clustering and characterization of repetitive
797sequences in next-generation sequencing data. *BMC Bioinformatics*. 2010;11:378–378.
79868. Novák P, Neumann P, Pech J, Steinhaisl J, MacAs J. RepeatExplorer: A Galaxy-based web
799server for genome-wide characterization of eukaryotic repetitive elements from next-generation
800sequence reads. *Bioinformatics*. 2013;29:792–3.
80169. Novák P, Ávila Robledillo L, Koblížková A, Vrbová I, Neumann P, Macas J. TAREAN: a
802computational tool for identification and characterization of satellite DNA from unassembled short
803reads. *Nucleic Acids Res*. 2017;45:e111–e111.
80470. Dierckxsens N, Mardulyn P, Smits G. NOVOPlasty: *de novo* assembly of organelle genomes
805from whole genome data. *Nucleic Acids Res*. 2016;gkw955.
80671. Chalopin D, Naville M, Plard F, Galiana D, Volff J-N. Comparative analysis of transposable
807elements highlights mobilome diversity and evolution in vertebrates. *Genome Biol Evol*.
8082015;7:567–80.
80972. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq
810data with DESeq2. *Genome Biol*. 2014;15:550.
81173. Shumate A, Zimin AV, Sherman RM, Puiu D, Wagner JM, Olson ND, et al. Assembly and
812annotation of an Ashkenazi human reference genome. *Genome Biol*. 2020;21:129.
81374. Tomaszekiewicz M, Medvedev P, Makova KD. Y and W chromosome assemblies: approaches
814and discoveries. *Trends Genet*. 2017;33:266–82.
81575. Ylla G, Nakamura T, Itoh T, Kajitani R, Toyoda A, Tomonari S, et al. Cricket genomes: the
816genomes of future food. *bioRxiv*. 2020; <http://biorxiv.org/lookup/doi/10.1101/2020.07.07.191841>

81776. Gregory TR. Chapter 1 - Genome Size Evolution in Animals. Burlington: Academic Press;
8182005. p. 3–87.
81977. Gregory TR. The Evolution of the Genome. In: Gregory TR, editor. Burlington: Academic
820Press; 2005.
82178. Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome.
822Nat Rev Genet. 2007;8:272–85.
82379. Sun YH, Xie LH, Zhuo X, Chen Q, Ghoneim D, Zhang B, et al. Domestic chickens activate a
824piRNA defense against avian leukosis virus. eLife. 2017;6:e24695.
82580. Vandeweghe MW, Platt RN, Ray DA, Hoffmann FG. Transposable element targeting by piRNAs
826in laurasiatherians with distinct transposable element histories. Genome Biol Evol. 2016;8:1327–37.
82781. Lim RSM, Kai T. A piece of the pi(e): The diverse roles of animal piRNAs and their PIWI
828partners. Sem Cell Develop Biol. 2015;47–48:17–31.
82982. Lower SS, McGurk MP, Clark AG, Barbash DA. Satellite DNA evolution: old ideas, new
830approaches. Curr Opin Genet Develop. 2018;49:70–8.
83183. Volpe TA. Regulation of heterochromatic silencing and Histone H3 Lysine-9 methylation by
832RNAi. Science. 2002;297:1833–7.
83384. White MJD, Dennis ES, Honeycutt RL, Contreras N, Peacock WJ. Cytogenetics of the
834parthenogenetic grasshopper *Warramaba virgo* and its bisexual relatives - IX. The ribosomal RNA
835cistrons. Chromosoma. 1982;85:181–99.
83685. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences
837of k-mers. Bioinformatics. 2011;27:764–70.
83886. Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simão FA, Ioannidis P, et al.
839OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant,
840archaeal, bacterial and viral orthologs. Nucleic Acids Res. 2017;45:D744–9.

84187. Suh A, Smeds L, Ellegren H. Abundant recent activity of retrovirus-like retrotransposons within
842and among flycatcher species implies a rich source of structural variation in songbird genomes. *Mol*
843*Ecol*. 2018;27:99–111.
84488. Katoh K, Toh H. Recent developments in the MAFFT multiple sequence alignment program.
845*Brief Bioinformatics*. 2008;9:286–98.
84689. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic
847genomes. *Mobile DNA*. 2015;6:11.
84890. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation
849sequencing data. *Bioinformatics*. 2012;28:3150–2.
85091. Song L, Florea L. Rcorrector: efficient and accurate error correction for Illumina RNA-seq
851reads. *GigaSci*. 2015;4:48.
85292. Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in
853metatranscriptomic data. *Bioinformatics*. 2012;28:3211–7.
85493. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*.
8552012;9:357–9.
85694. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor:
857open software development for computational biology and bioinformatics. *Genome Biol*. 2004;16.
85895. R Core Team. R: a language and environment for statistical computing. Available from: [https://](https://www.R-project.org/)
859www.R-project.org/
86096. Gordon A, Hannon GJ. Fastx-toolkit. FASTQ/A short-reads pre-processing tools. Available
861from: http://hannonlab.cshl.edu/fastx_toolkit
86297. Junier T, Pagni M. Dotlet: diagonal plots in a web browser. *Bioinformatics*. 2000;16:178–9.
86398. Seibt KM, Schmidt T, Heitkam T. FlexiDot: highly customizable, ambiguity-aware dotplots for
864visual sequence analyses. *Bioinformatics*. 2018;34:3575–7.

86599. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: Molecular
866evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum
867parsimony methods. *Mol Biol Evol.* 2011;28:2731–9.
868100. Frydrychová R, Grossmann P, Trubac P, Vítková M, Marec F. Phylogenetic distribution of
869TTAGG telomeric repeats in insects. *Genome.* 2004;47:163–78.

870**Table 1: Pseudohaploid assembly statistics across the four chromosomal races of the *viatica* species group.**

	P24X0 male	P24XY male	P45bX0 male	P45bXY male	
Total millions of reads (2x150 bp)	1,883	1,582.	1,759	1,577	10X Genomics Chromium linked reads
Molecule length (kb)	41.18	17.44	56.11	30.44	Average input molecule lengths
Long scaffold (kb)	44.43	100.68	88.04	99.29	Number of scaffolds >= 10 kb
Scaffold N50 (kb)	316.69	34.85	51.83	36.43	N50 scaffold size
Edge N50 (kb)	8.44	6.24	8.24	6.73	N50 edge size
Contig N50 (kb)	35.69	26.94	33.34	29.11	N50 contig size
Phaseblock N50 (kb)	651.54	50.13	115.56	53.56	N50 phase block size
Assembly GC content %	37.99	37.93	38.09	38.01	
Assembled genome size (Gb)	3.02	2.73	3.27	2.94	Only scaffolds >= 10 kb
Genome size estimation (Gb)	3.30	3.36	3.52	3.82	Supernova (computed from k-mer distributions)
	5.42	5.65	5.76	6.32	FindGSE (computed from k-mer distributions)
Average genome size (Gb)	4.26	4.50	4.64	5.07	Based on k-mer distributions

Table 2. Assembly-based quantification of repeats. Copy number, total base pair and density of different classes of repeats annotated by RepeatMasker using a combined library of the RepeatModeler *de-novo* library from each race (manual curation of the P24X0 library, used for re-classification of the three other *de-novo* libraries), the Arthropod library from Repbase, and libraries from HelitronScanner and RepeatExplorer2, across males of four chromosomal races of the *viatica* species group. TR = tandem repeats. GP = genome proportion (% assembly).

Repeat type	P24X0			P24XY			P45bX0			P45bXY		
	Copies	Total bp	GP	Copies	Total bp	GP	Copies	Total bp	GP	Copies	Total bp	GP
SINE	355,321	57,777,060	1.45	476,672	84,209,572	2.31	498,884	86,588,794	2.16	331,980	55,473,057	1.45
LINE	1,630,136	751,701,393	18.88	1,544,315	715,290,166	19.6	1,603,310	792,142,166	19.79	1,780,161	819,481,256	21.40
LTR	270,520	191,052,819	4.80	298,156	160,264,290	4.39	316,814	207,702,027	5.19	303,192	169,275,629	4.42
DNA	2,269,382	712,386,285	17.89	2,136,964	678,964,713	18.61	2,179,781	734,004,927	18.34	2,165,693	714,794,492	18.66
RC	3,194,220	663,607,802	16.72	8,654,890	708,282,512	19.45	3,887,696	878,010,926	21.88	3,569,564	782,636,309	20.45
TR	5,269	790,157	0.02	3,294	506,934	0.01	10,988	1,403,691	0.04	232	12,409	0.01
Unclassified	889,785	237,465,164	5.98	795,764	240,830,299	6.62	946,725	290,976,813	7.29	8,992,548	237,984,790	6.24
Total	8,614,633	2,614,780,680	65.74	13,910,055	2,588,348,486	70.99	9,444,198	2,989,425,653	74.69	17,143,370	2,779,657,942	72.62

Table 3. Read-based quantification of repeats by sampling 4 million of read pairs per library. Copy number, total base pair and density of different classes of TEs by RepeatMasker using a combined library of the RepeatModeler *de-novo* library from each race (manual curation of the P24X0 library, used for re-classification of the three other de-novo libraries), the Arthropod library from Repbase, and libraries from HelitronScanner and RepeatExplorer2, across males of four chromosomal races of the *viatica* species group. TR = tandem repeats. GP = genome proportion (% reads).

Repeat type	P24X0 GP	P24XY GP	P45bX0 GP	P45bXY GP
SINE	1.23	1.49	1.81	1.27
LINE	15.48	14.26	14.91	15.09
LTR	3.39	3.29	3.81	3.28
DNA	13.83	14.43	13.95	13.77
RC	23.17	24.84	28.73	25.67
TR	7.82	6.06	7.53	8.85
Unclassified	9.86	8.65	8.50	7.51
Total	74.78	73.02	79.44	75.44

877

Table 4. Differentially expressed TE subfamilies containing only recent TEs (genomic copies with < 5 % K2P distance from the consensus sequences) between sexes and tissues in two chromosomal races of the *viatica* species group.

Race	Sample	No. of expressed TEs	SBTEs (%)	MBTEs (%)	FBTEs (%)
P24X0	Head	574	23 (4.05)	21 (3.66)	2 (0.35)
	Gonad	598	152 (25.42)	62 (10.37)	90 (15.05)
	Leg	1,362	4 (0.29)	3 (0.23)	1 (0.07)
P24XY	Head	1,415	24 (1.71)	21 (1.48)	3 (0.21)
	Gonad	1,415	117 (8.27)	69 (4.88)	48 (3.39)
	Leg	775	4 (0.52)	0	4 (0.52)

We report the total number of TEs expressed, and those TEs that were sex-biased with $\log_2\text{FC} > 0$ and $\log_2\text{FC} < 0$ and adjusted $P < 0.05$; $\log_2\text{FC} = \log_2$ fold-change. SBTEs = sex-biased TEs; MBTE = male-biased TEs.

Table 5: Average of abundance and divergence of the 10 satDNAs shown in the Figure 6.

Repeat	Abundance (%)				Divergence (% K2P)				Differential copy amplification (Kruskal-Wallis test)
	P24X0	P24XY	P45bX0	P45bXY	P24X0	P24XY	P45bX0	P45bXY	
satDNA-1	1.48	0.13	0.23	0.06	13.94	19.60	20.28	19.86	$P < 0.03$
satDNA-2	0.96	0.01	0.01	0.02	7.12	9.50	11.31	10.18	$P < 0.01$
satDNA-8	0.15	0.01	0.06	0.08	5.10	4.72	6.57	4.01	$P < 0.01$
satDNA-17	0.06	0.50	0.57	0.75	24.4	15.77	19.48	19.90	$P < 0.01$
satDNA-27	0.02	0.03	0.03	0.19	6.73	9.51	8.44	9.29	$P < 0.01$
satDNA-45	0.39	0.53	0.32	0.51	25.33	17.38	23.58	25.53	$P < 0.01$
satDNA-46	0.01	0.51	0.05	0.44	16.64	16.42	17.12	18.30	$P > 0.33$
satDNA-47	0.01	0.27	0.03	0.01	11.57	10.01	9.83	11.14	$P < 0.01$
satDNA-90	1.20	1.05	1.20	1.20	22.82	23.02	23.85	23.67	$P < 0.01$
satDNA-111	0.004	0.01	0.005	0.02	22.46	15.34	22.84	9.78	$P > 0.81$

Figure 1. Overview of the karyotype, BUSCO assembly assessment, and repeat composition across four chromosomal races of the *viatica* species group. **a)** Karyotypes of the four chromosomal races. Chromosomes are aligned based on centromere position. Only haploid sets of chromosomes in males are shown. Two independent emergences of neo-XY sex chromosomes via fusions between the ancestral X chromosome (red) and one of the autosomes (black) are highlighted. **b)** BUSCO completeness assessment for genome assembly quality control. Color-coded bar plot showing the proportions of BUSCO genes classified as complete, complete single-copy, complete duplicated, fragmented, and missing. **c-d)** Color-coded bar plot illustrating the proportions of major TE groups, unclassified repeats, tandem repeats (TRs), and non-repetitive regions detected in the assembled genomes (**d**) and in the sequenced reads (**c**). The major TE groups included manually curated RepeatModeler consensus sequences and Helitrons (RC, rolling circle replication) identified by HelitronScanner.

Figure 2. Helitron structures found in the genome assemblies of the *viatica* species group. a) Typical autonomous animal Helitron. b) Non-autonomous animal Helitron. (c) The new identified Helitron family identified in the P24X0 chromosomal race, named Tukutron. The new Helitron bears entire and truncated copies of short interspersed elements (SINEs) which are similar to SINEs in the *Locusta migratoria* genome (SINE2-3_Lmi in Repbase).

Figure 3. Repeat landscapes illustrating recent accumulation of 49 superfamilies of transposable elements across four chromosomal races of the viatica species group. a) P24X0. b) P24XY. c) P45bX0. d) P45bXY. Color-coded bar plots were generated from RepeatMasker .align files after selection of all TE copies with K2P distances < 5 %. The graphs represent the base pairs occupied of a given TE superfamily (y axis) in the different genomes analyzed, binned according to K2P distances to their corresponding consensus sequence (x axis, K2P distance from 0 to 5%). We considered copies in the divergence bins < 5 % as recent TEs, likely corresponding to copies of recently active elements. Plots showing all divergence bins (K2P distance from 0 to 50 %) are shown in Additional file 2: Figure S1.

Figure 5. Repeat landscapes of the main groups of tandem repeats (TRs) detected in the sequenced reads of the four chromosomal races of the *viatica* species group. a) P24X0. b) P24XY. c) P45bX0. d) P45bXY. Color-coded bar plots are based on RepeatMasker showing K2P distance to their corresponding consensus sequence on the x axis and the TR abundance on the y axis. Color-coded bar plots show the proportions of TRs classified as satellite (45-81 families), telomere repeat and the multigene families U snRNA genes (U1, U2, U5, U6), histone genes (H1, H2A, H2B, H3, H4), 45S rRNA and 5S rRNA genes. More detailed repeat landscapes of all TR families are shown in the Additional file 2: Figure S2.

Figure 6. satDNA presence and divergence across four chromosomal races of the *viatica* species group. **a)** Venn diagram showing the number of satDNA families shared among the chromosomal races. **b)** Repeat landscapes based on RepeatMasker illustrating the abundance and divergence of 10 randomly selected satDNAs. **a)** P24X0, **b)** P24XY, **c)** P45bX0 and **d)** P45bXY. Color-coded bar plots showing the genome proportion (*y* axis) for each type of satDNAs in the sequenced reads analyzed, binned according to K2P distances to their corresponding consensus sequence (*x* axis, K2P value from 0 to 50 %). Copies in low-divergence bins are very similar to the consensus sequence and likely correspond to very recent copies. Out of the 102 satDNA families shared between all four races, there were 50 satDNA families showing differential amplification in copy number (*cv* > 80 %; Additional file 2: Table S8). From these 50 families, 10 randomly selected satDNAs are shown here for visualization of differential proliferation of satDNAs.