













RESEARCH ARTICLE

First draft genome assembly of the desert locust, *Schistocerca gregaria* [version 1; peer review: awaiting peer review]

Heleen Verlinden ¹, Lieven Sterck ^{2,3}, Jia Li^{2,3}, Zhen Li ^{2,3}, Anna Yssel⁴, Yannick Gansemans ^{5,6}, Rik Verdonck^{1,7}, Michiel Holtorf¹, Hojun Song ⁸, Spencer T. Behmer⁸, Gregory A. Sword ⁸, Tom Matheson ⁹, Swidbert R. Ott⁹, Dieter Deforce ^{5,6}, Filip Van Nieuwerburgh ^{5,6}, Yves Van de Peer²⁻⁴, Jozef Vanden Broeck ¹

¹Laboratory of Molecular Developmental Physiology and Signal Transduction, KU Leuven, Leuven, 3000, Belgium

²Laboratory of Bioinformatics and Evolutionary Genomics, Ghent University, Ghent, 9000, Belgium

³Center for Plant Systems Biology, Ghent University - VIB, Ghent, 9052, Belgium

⁴Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria, 0002, South Africa

⁵Laboratory of Pharmaceutical Biotechnology, Ghent University, Ghent, 9000, Belgium

⁶NXTGNT, Ghent University, Ghent, 9000, Belgium

⁷Station d' Ecologie Théorique et Expérimentale, UMR 5321 CNRS et Université Paul Sabatier, Moulis, 09200, France

⁸Department of Entomology, Texas A&M University, College Station, Texas, TX 77843-2475, USA

⁹Department of Neuroscience, Physiology and Behaviour, University of Leicester, Leicester, LE1 7RH, UK

V1 **First published:** 27 Jul 2020, 9:775
<https://doi.org/10.12688/f1000research.25148.1>
Latest published: 27 Jul 2020, 9:775
<https://doi.org/10.12688/f1000research.25148.1>

Abstract

Background: At the time of publication, the most devastating desert locust crisis in decades is affecting East Africa, the Arabian Peninsula and South-West Asia. The situation is extremely alarming in East Africa, where Kenya, Ethiopia and Somalia face an unprecedented threat to food security and livelihoods. Most of the time, however, locusts do not occur in swarms, but live as relatively harmless solitary insects. The phenotypically distinct solitary and gregarious locust phases differ markedly in many aspects of behaviour, physiology and morphology, making them an excellent model to study how environmental factors shape behaviour and development. A better understanding of the extreme phenotypic plasticity in desert locusts will offer new, more environmentally sustainable ways of fighting devastating swarms.

Methods: High molecular weight DNA derived from two adult males was used for Mate Pair and Paired End Illumina sequencing and PacBio sequencing. A reliable reference genome of *Schistocerca gregaria* was assembled using the ABySS pipeline, scaffolding was improved using LINKS.

Results: In total, 1,316 Gb Illumina reads and 112 Gb PacBio reads were produced and assembled. The resulting draft genome consists of 8,817,834,205 bp organised in 955,015 scaffolds with an N50 of 157,705 bp, making the desert locust genome the largest insect genome sequenced and assembled to date. In total, 18,815 protein-encoding genes are

Open Peer Review

Reviewer Status AWAITING PEER REVIEW

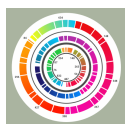
Any reports and responses or comments on the article can be found at the end of the article.

predicted in the desert locust genome, of which 13,646 (72.53%) obtained at least one functional assignment based on similarity to known proteins.

Conclusions: The desert locust genome data will contribute greatly to studies of phenotypic plasticity, physiology, neurobiology, molecular ecology, evolutionary genetics and comparative genomics, and will promote the desert locust's use as a model system. The data will also facilitate the development of novel, more sustainable strategies for preventing or combating swarms of these infamous insects.

Keywords

Eco-devo, large genome size, locust plague, Orthoptera, pest insect, phenotypic plasticity, polyphenism, swarm



This article is included in the **Draft Genomes** collection.

Corresponding authors: Yves Van de Peer (yypee@psb.vib-ugent.be), Jozef Vanden Broeck (jozef.vandenbroeck@kuleuven.be)

Author roles: **Verlinden H:** Conceptualization, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Visualization, Writing – Original Draft Preparation; **Sterck L:** Conceptualization, Data Curation, Formal Analysis, Methodology, Software, Supervision, Visualization, Writing – Original Draft Preparation; **Li J:** Data Curation, Formal Analysis, Methodology, Software, Visualization, Writing – Original Draft Preparation; **Li Z:** Data Curation, Formal Analysis, Methodology, Software, Visualization, Writing – Original Draft Preparation; **Yssel A:** Data Curation, Formal Analysis, Writing – Review & Editing; **Gansemans Y:** Data Curation, Formal Analysis, Methodology, Software, Writing – Original Draft Preparation; **Verdonck R:** Data Curation, Formal Analysis, Investigation, Visualization, Writing – Review & Editing; **Holtorf M:** Investigation, Writing – Review & Editing; **Song H:** Conceptualization, Funding Acquisition, Writing – Review & Editing; **Behmer ST:** Conceptualization, Funding Acquisition, Writing – Review & Editing; **Sword GA:** Conceptualization, Funding Acquisition, Writing – Review & Editing; **Matheson T:** Conceptualization, Funding Acquisition, Writing – Review & Editing; **Ott SR:** Conceptualization, Funding Acquisition, Writing – Review & Editing; **Deforce D:** Resources, Writing – Review & Editing; **Van Nieuwerburgh F:** Conceptualization, Resources, Supervision, Writing – Review & Editing; **Van de Peer Y:** Conceptualization, Funding Acquisition, Project Administration, Resources, Supervision, Writing – Review & Editing; **Vanden Broeck J:** Conceptualization, Funding Acquisition, Project Administration, Resources, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the Special Research Fund of KU Leuven (BOF grant C14/15/050 to JvdB and HV), the Research Foundation of Flanders (FWO grants: postdoctoral fellowship 64322 to HV, G0F2417N to JvdB, G090919N to JvdB and YVdP); the Special Research Fund of Ghent University (BOFPDO2018001701 to ZL), the Department of Research and Innovation of the University of Pretoria (grant A0C827 to AY); the U.S. National Science Foundation (IOS-1253493 and IOS-1636632 to HS), the U.S. Department of Agriculture (hatch grant TEX0-1-6584 to HS) and the Biotechnology and Biological Sciences Research Council UK (BBSRC; research grant BB/L02389X/1 to SRO and TM).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2020 Verlinden H *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Verlinden H, Sterck L, Li J *et al.* **First draft genome assembly of the desert locust, *Schistocerca gregaria* [version 1; peer review: awaiting peer review]** F1000Research 2020, 9:775 <https://doi.org/10.12688/f1000research.25148.1>

First published: 27 Jul 2020, 9:775 <https://doi.org/10.12688/f1000research.25148.1>

Introduction

Locust plagues have been recorded since Pharaonic times in ancient Egypt. In the Bible (*Exodus 10*), locust swarms are described as one of the major destructive plagues and still today they form a serious threat to crops and food security of over 60 countries across more than 20% of the world’s total land surface (Figure 1a). Swarms can cover areas up to several hundred square kilometres and migrate up to 200 km per day. Per square kilometre, a swarm that contains about 40 million locusts can

eat the same amount of food in one day as about 35,000 people. The damage done by a locust plague is on the same level as a major drought (FAO Locust Watch; De Vrejer *et al.*, 2012). The long-term socio-economic impact of these swarms is significant. The loss of harvest is disastrous for local farmers and leads to towering local food prices, also affecting non-farming families. The poorest households are often hit the hardest. Malnourishment of children and expecting mothers endangers their long-term health and growth. School enrolment

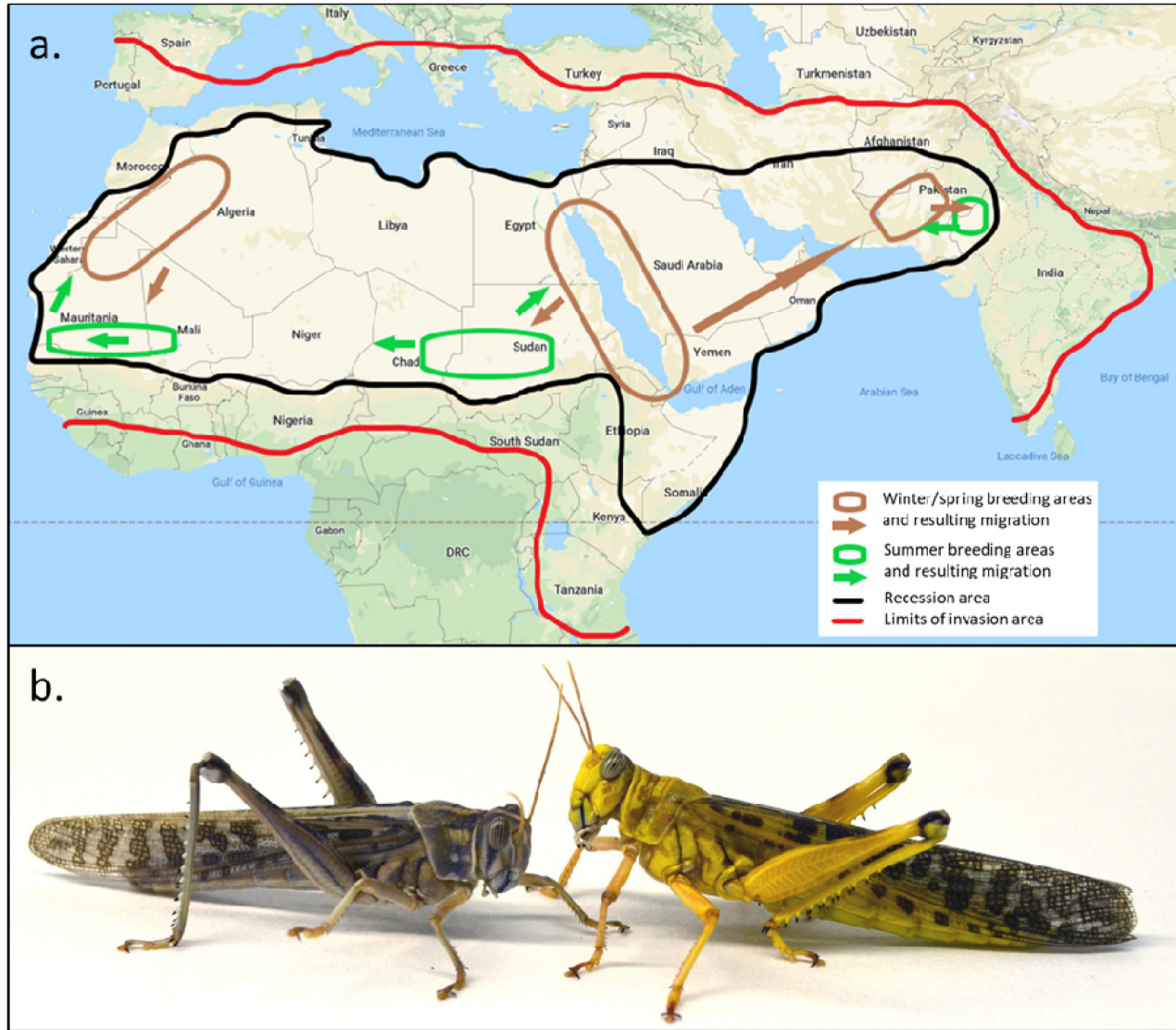


Figure 1. Geographical distribution of the desert locust and a picture of two adult male desert locusts, one in the solitary phase and the other in the gregarious phase. (a) Geographic distribution of the desert locust. During ‘recession’ periods, desert locusts are restricted to the semi-arid and arid regions of Africa, the Arabian Peninsula and South-West Asia that receive less than 200 mm of annual rain. The recession area covers about 16 million km² in 30 countries. Within this recession area, locusts move seasonally between winter/spring and summer breeding areas. During outbreaks, desert locusts may spill into more fertile adjacent regions, threatening an area of some 29 million km² comprising 60 countries as outbreaks escalate into upsurges and further into plagues. The recession breeding areas and migration patterns may have predictive value to understand how the swarms will migrate. Figure based on information from FAO Locust Watch, map derived from Google Map Data ©2020 Google. **(b)** Phase polyphenism in desert locusts, using the example of sexually mature males. The gregarious male (right) has bright warning colours to avoid predation, while the solitary male relies on camouflage colours. In this staged scene, the solitary male was forced into close proximity of the gregarious male and is seen retreating from its conspecific. Photo by H. Verlinden and R. Verdonck.

rate fell by a quarter during plagues in 1987–89 in Mali, with girls being particularly affected (Courcoux, 2012). Human activities in turn affect the propensity of locusts to swarm through factors such as land use (e.g. agriculture, wood extraction, urbanization), political relations between affected countries and the effects of climate change (FAO Locust Watch, <http://www.fao.org/ag/locusts/en/info/info/index.html>; Cullen *et al.*, 2017; Meynard *et al.*, 2020).

Desert locusts (*Schistocerca gregaria* Forskål) are grasshoppers (Orthoptera: Acrididae) that exhibit ‘phase polyphenism’, an extreme form of phenotypic plasticity that evolved as an adaptation to the drastic changes that can occur in their environment. Locusts can develop into two extremely divergent, population density-dependent phenotypes, which are tailored to very different ecological requirements. Under low population densities, locusts appear in the solitary phase and live a solitary life in which they avoid each other. In periods with abundant rainfall, rapid vegetation growth creates a favourable habitat that permits large increases in local population sizes. However, when food becomes scarce again, solitary locusts are forced to aggregate on the remaining plants. This crowding causes the transformation into the swarming gregarious phase, beginning with rapid changes in behaviour that include a switch to increased locomotion and mutual attraction. The prolonged crowding drives slower but equally profound changes in colouration, morphology (Figure 1b) and physiology. Compounded across multiple generations, locust populations can aggregate further into huge, ruinous swarms capable of crossing continents and oceans in search of food. Populations may crash in the absence of sufficient resources or following human intervention, leading once more to scattered low density solitary phase populations. The transition between locust phases is thus reversible and occurs gradually through the expression of intermediate phenotypic states (Cullen *et al.*, 2017; Pener & Simpson, 2009; Symmons & Cressman, 2001; Verlinden *et al.*, 2009).

Orthoptera (grasshoppers, crickets and allies) belong to the Polyneoptera, a clade that represents one of the major lineages of winged insects (Pterygota) and comprises around 40,000 known species and ten orders of hemimetabolous insects (Misof *et al.*, 2014; Wipfler *et al.*, 2019). Other major neopteran (Pterygota that can flex their wings over their abdomen) lineages are Acercaria (mostly sucking insects such as lice or true bugs) and Holometabola (insects with complete metamorphosis). At present, only 25 sequenced polyneopteran genomes are reported on NCBI and i5k (http://i5k.github.io/arthropod_genomes_at_ncbi), unequally distributed over five different orders (Extended data, Supplementary Table S1 (Verlinden *et al.*, 2020)). When including *S. gregaria*, the genomes of five orthopteran species, representing five different subfamilies, are now available. In addition to representing a paradigmatic example of phenotypic plasticity, the desert locust is an important research model for generating advances in a wide variety of fundamental and applied scientific areas, including biomechanics, ecology, pest control, neurobiology and physiology. For instance, the relatively large body size of locusts has been

instrumental in discovery of a multitude of insect neuropeptides (Schoofs *et al.*, 1997). Moreover, the globally increasing interest in the use of insects as food or feed also applies to the desert locust, which is a highly nutrient-rich, edible insect that is gaining much attention as a potential, climate-friendly food source (van Huis *et al.*, 2013).

The devastating socio-economic impact of locust swarms, together with the opportunity this species offers to investigate the phenotypic interface of molecular processes and environmental cues highlight the importance of sequencing the desert locust genome. However, the extremely large estimated genome size of 8.55 Gb (Camacho *et al.*, 2015; Fox, 1970; John & Hewitt, 1966; Wilmore & Brown, 1975) predicted a formidable challenge. Moreover, previous transcriptomics and chromosome size data from the desert locust (Badisco *et al.*, 2011; Camacho *et al.*, 2015), as well as comparisons with the genome of the distantly related migratory locust, *Locusta migratoria* (6.5 Gb; Wang *et al.*, 2014), suggested that the non-coding part of the desert locust genome might be greatly expanded as compared to other insect genomes, presenting additional challenges to sequencing and assembly. Our team has overcome these hurdles and presents here the ~8.8 Gb genome of the desert locust assembled from short Illumina Mate Pair (MP) and Paired End (PE) reads and long PacBio reads. This new genomic resource, the largest insect genome yet sequenced and assembled, will complement decades of research on this species, enhancing the desert locust’s role as an important comparative model system. The genome will permit exciting new opportunities to examine mechanisms of phenotypic plasticity, social behaviour, physiological and morphological specialization. Moreover, it will open up new avenues to find better ways of fighting the notorious swarms they can cause. The desert locust genome will also enable better understanding of genome size evolution and the early phylogeny of winged insects.

Methods

Sequencing strategy

A hybrid sequencing approach was adopted consisting of both Illumina short read sequencing to get sufficient coverage for accurate contig assembly, and complementary PacBio long read sequencing to allow efficient scaffolding of the contig assembly. The Illumina and first PacBio sequencing were performed on high-molecular-weight DNA derived from the central nervous system (central brain, optic lobes, ventral nerve cord), fat body and testes of one adult male inbred for seven generations. A second round of PacBio sequencing used DNA from another male from the same lineage, with two additional generations of inbreeding (for details on the animal material and genomic DNA extraction, see Extended data, Supplementary Methods (Verlinden *et al.*, 2020)).

Illumina sequencing

The concentration of the *S. gregaria* high molecular weight DNA sample was measured with PicoGreen (Invitrogen) fluorimetry, after which DNA integrity was confirmed by gel electrophoresis (1% E-Gel; Invitrogen). The sample was divided for Illumina MP and PE sequencing library preparation.

The MP sequencing library was prepared from 1 µg of the sample with a “Nextera Mate Pair Library prep kit” (Illumina). The PE library was prepared with a “NEBNext Ultra II library prep kit” (NEB) from 2 µg of the sample, sheared to 500 bp fragments using an S2 focused-ultrasonicator (Covaris). Size selection (600–700 bp) was performed for both libraries in a 2% E-Gel (Invitrogen). The quality of the libraries was confirmed with a Bioanalyzer High Sensitivity DNA Kit (Agilent). The MP and PE libraries were quantified by qPCR, according to Illumina’s “Sequencing Library qPCR Quantification protocol guide” (version February 2011) and pooled at a molar ratio of 25% MP – 75% PE for sequencing on Hiseq3000 (2 × 150 cycles, 16 lanes; Illumina).

PacBio sequencing

The library preparation for PacBio sequencing was performed with a “SMRTbell Template Prep Kit 1.0” according to the PacBio protocol (version 100-286-000). For each of the two libraries, 10 µg of the *S. gregaria* high-molecular-weight DNA was used as input in two parallel 50-µl reactions.

For library size selection, a “0.75% Dye-Free Agarose Gel Cassette” (ref: BLF7510) was used on a Blue Pippin (Sage Science) with the “0.75% DF Marker S1 high-pass 15–20kb” protocol for a lower cut-off of 12 kb. Fragment size distribution was determined with a “DNA 12000 kit” (ref: 5067-1508) for the first library and a “Fragment Analyzer (Agilent) - High Sensitivity Large Fragment 50 kb kit” (ref: DNF-464-0500) for the second library. The resulting libraries had an average length of 16.5 and 22 kb, respectively.

No extension time was used for the sequencing as recommended for size selected libraries in the “Quick Reference Card 101-461-600 version 07”. The first run was performed on a PacBio RSII System (V4.0 chemistry, polymerase P6). Fifteen additional runs were performed on a PacBio Sequel system with 2.0 Chemistry, polymerase and SMRTCells. The same conditions were used to sequence 20 more SMRTCells with the second library on the PacBio Sequel system.

Genome assembly

PE short read data were pre-processed with bbduk v38.20 from the BBTools package to remove adapters and low-quality reads. Illumina MP read data were cleaned and separated into true MP data and likely MP data in nxTrim (O’Connell *et al.*, 2015). The long-read PacBio data were pre-processed using CANU v1.7 (Koren *et al.*, 2017) to obtain trimmed and corrected reads. Cleaned short-read PE and MP data were then assembled using the ABySS v2.1.1 pipeline (Simpson *et al.*, 2009) up to scaffold stage, using a k-mer value of 120. Parameters for ABySS were optimized away from default values to achieve better performance (for all parameter settings see *Extended data*, Supplementary Table S2 (Verlinden *et al.*, 2020)). The assembly was further improved by using the PacBio data as input for LINKS (Warren *et al.*, 2015).

Annotation of repetitive elements and noncoding RNAs

Two strategies were used to identify and annotate repetitive elements. First, *de novo* annotation was carried out by

RepeatModeler v2.0 and LTR_FINDER v1.0.7 (Xu & Wang, 2007) to build a custom repeat library. Second, a homology-based approach was used to search for repetitive elements in the assembled genome using the repetitive element library of RepeatMasker v4.1.0 and RepeatProteinMask v4.1.0. The results of both strategies were combined into a non-redundant set of repetitive elements. Subsequently, the library was used to mask repetitive elements by employing RepeatMasker v4.1.0 (Tarailo-Graovac & Chen, 2009).

Transfer RNAs (tRNAs) were predicted by tRNAscan-SE v1.31 (Lowe & Eddy, 1997) with default parameters. To predict non-coding RNAs (ncRNAs), such as microRNAs (miRNAs), small nuclear RNAs (snRNAs), and ribosomal RNAs (rRNAs), the desert locust genome was screened against the RNA families (Rfam) v14.1 database (Griffiths-Jones *et al.*, 2003) by the cmscan program of Infernal v1.1.2 (Nawrocki & Eddy, 2013). To supplement our predictions of miRNAs, miRNA sequences from the *L. migratoria* genome (Wang *et al.*, 2015) were extracted and searched in the *S. gregaria* genome by BLASTN with options “-task blastn-short -ungapped -penalty -1 -reward 1” (Camacho *et al.*, 2008). The alignment result was filtered using a mismatch cutoff of 3 bp. Specifically, the stem-loop structure of each potential miRNA was predicted by miRNAfold (Tav *et al.*, 2016) using each alignment with 110 bp upstream and downstream sequences. Then the RNAfold program of ViennaRNA v2.4.14 (Lorenz *et al.*, 2011) was used to calculate the minimum free energy (MFE) of each stem-loop structure. If a potential miRNA had several predicted stem-loop structures, the one with the minimum MFE was selected as representative. Putative miRNAs located within protein coding sequences or repetitive elements were discarded. Finally, the results based on Rfam and the migratory locust genome were combined into a non-redundant prediction of miRNAs.

Gene prediction and functional annotation

Protein-coding genes in the desert locust genome were predicted using three approaches. (1) RNA-Seq reads (see *Extended data*, Supplementary Methods (Verlinden *et al.*, 2020)) were mapped to the desert locust genome using HISAT2 v2.1.0 (Kim *et al.*, 2015) with parameter “--max-intronlen” set to 1,000,000 to increase the maximum allowed intron length during read mapping. Then, StringTie v2.1.1 (Pertea *et al.*, 2015) was used to assemble potential transcripts based on RNA-Seq alignments to the desert locust genome. Subsequently, TransDecoder v5.0.2 was used to identify open reading frames (ORFs) within the assembled transcripts which resulted in 20,201 ORFs with start and/or stop codons. We also built *de novo* assembled transcripts based on the pooled RNA-Seq reads of all samples with Trinity v2.8.4 (Grabherr *et al.*, 2011; Haas *et al.*, 2013) and obtained 285,499 transcripts (including isoforms), of which 57,870 putative protein-coding transcripts and 305 rRNA candidates were identified by Trinotate v3.1.1 (Bryant *et al.*, 2017). This was complemented with 34,974 ESTs of the desert locust from NCBI (Badisco *et al.*, 2011). The assembled transcripts and ESTs were then aligned to the desert locust genome with Program to Assemble Spliced Alignments (PASA v2.4.1) (Haas *et al.*, 2003). (2) For *ab initio* gene prediction, we used a hard-masked genome in which genomic repetitive

elements were substituted by 'N'. To build a training set for the *ab initio* gene predictors, we extracted 498 complete genes with both start and stop codons from the 500 longest ORFs predicted by TransDecoder, based on the above RNA-Seq analysis with HISAT2 and StringTie. Augustus v3.3.3 (Stanke *et al.*, 2006) SNAP v2006-07-28 (Korf, 2004) and GlimmerHMM v3.0.4 (Majoros *et al.*, 2004) were trained on this training set and then used to predict potential gene models. Furthermore, combined with the RNA-Seq alignments, BRAKER2 v2.1.5 (Hoff *et al.*, 2019) was used to predict protein-coding genes based on the above-mentioned training model of Augustus. (3) The proteomes of the migratory locust, *Locusta migratoria* (Wang *et al.*, 2014); the African malaria mosquito, *Anopheles gambiae*; the domestic silk moth, *Bombyx mori*; the fruit fly, *Drosophila melanogaster*; the kissing bug, *Rhodnius prolixus*; the red imported fire ant, *Solenopsis invicta*; the red flour beetle, *Tribolium castaneum*; and the Nevada dampwood termite, *Zootermopsis nevadensis* from Ensembl Metazoa (release-47), as well as the proteins in UniRef100 (release-2020_01) for the clade Polyneoptera (Taxonomy ID: 33341) were used to assist gene predictions with homologous proteins. Exonerate v2.4.0 (Slater & Birney, 2005) was used to perform spliced alignments of the proteins with the maximum intron length set to 1 Mb. To integrate the predictions from all three gene-prediction approaches, EvidenceModeler v1.1.1 (Haas *et al.*, 2008) was used to produce a non-redundant gene set. Functional annotation of the predicted

protein-coding genes was done by running BlastP (Altschul *et al.*, 1990) using an e-value cut-off of 1×10^{-5} against the public protein databases Uniprot/SwissProt (Magrane, 2011; The UniProt Consortium, 2019) and NCBI NR (RefSeq non-redundant protein record). Protein family (Pfam) domain information and Gene Ontology (GO) terms were added using InterProScan (Mitchell *et al.*, 2019).

Results and discussion

Genome size and assembly

Initial input data for the assembly comprised (i) 1,316 Gb of Illumina short read data, of which 1,009 Gb remained after cleaning and trimming, and (ii) 112 Gb of long reads from PacBio sequencing. The resulting assembly, using the ABySS pipeline, consisted of 8.5 Gb in ~1.6 M contigs with an N50 of 12,027 bp. Scaffolding with the MP data using ABySS resulted in 8.6 Gb in 1.2 M scaffolds with an N50 of 66,194 bp. The PacBio data as input for LINKS further improved the scaffolded assembly derived from ABySS, doubling the N50 and maximum length and reducing the number of sequences by half. The final assembly consists of 8,817,834,205 bp organised in 955,015 scaffolds with an N50 of 157,705 bp (Table 1).

Repetitive elements and noncoding RNAs

In total, repetitive elements account for 62.55% of the desert locust genome (Table 2), which is more than the 58.86% repetitive

Table 1. Results of the assembly for the desert locust genome.

	Total	Total size (bp)	N50 (bp)	N90 (bp)	Largest (bp)	Mean length (bp)
Contigs	1,648,200	8,561,922,307	12,027	5,375	202,979	5,194.71
Scaffolds (MP)	1,233,802	8,632,364,377	66,194	15,575	1,561,787	8,350.11
Scaffolds (PacBio)	955,015	8,817,834,205	157,705	29,453	3,339,430	9,233.20

Scaffolds (MP), Scaffolds reached with the Mate Pair data using the ABySS pipeline; Scaffolds (PacBio), improved scaffolds with the PacBio data as input for LINKS; N50, the sequence length of the shortest contig/scaffold at 50% of the total genome length; N90, the sequence length of the shortest contig/scaffold at 90% of the total genome length

Table 2. Repetitive elements in the genomes of the desert locust, *Schistocerca gregaria*, and the migratory locust, *Locusta migratoria* (Wang *et al.*, 2014).

Repeat Types	<i>Schistocerca gregaria</i>		<i>Locusta migratoria</i>	
	Length (bp)	P%	Length (bp)	P%
DNA	2,390,333,660	27.1	1,480,538,225	22.69
LINE	2,438,094,307	27.6	1,332,720,207	20.42
SINE	28,032,199	0.32	141,176,698	2.16
LTR	637,406,118	7.23	508,675,263	7.80
Other	165	0.00	32,017	0.00
Unknown	871,233,596	9.88	406,097,360	6.22
Total	5,515,243,572	62.55	3,840,808,141	58.86

DNA, DNA transposons; LINE, long interspersed nuclear element retrotransposon; SINE, short interspersed nuclear element retrotransposon; LTR, long terminal repeat retrotransposon; Other, repeats classified to other than the above mentioned types; Unknown, repeats that cannot be classified; P%, percentage of the genome.

elements in the published migratory locust genome (Wang *et al.*, 2014). Screening the desert locust genome against the Rfam v14.1 database identified 121,581 tRNAs, 1,302 rRNAs, 121 miRNAs, and 361 snRNAs (Extended data, Supplementary Table S3 (Verlinden *et al.*, 2020)).

In addition to the 121 evolutionary conserved miRNAs identified from Rfam, blasting with miRNAs previously identified in the migratory locust (from small RNA sequencing-based and homology-based approaches; Wang *et al.*, 2015) identified a further 686 miRNAs in the desert locust genome, resulting in a total of 807 identified miRNAs (Extended data, Supplementary Table S3 (Verlinden *et al.*, 2020)). Of these 807 miRNAs, 676 are located on short scaffolds without any protein-coding gene. Among the 121 miRNAs identified based on Rfam, 81 have no homologs in the migratory locust genome.

Protein-coding genes

In total, 18,815 protein-encoding genes are predicted in the desert locust genome (Extended data, Supplementary Table S4 (Verlinden *et al.*, 2020)). The average pre-mRNA length is 54,426 bp, with an average coding sequence (CDS) length of 1,137 bp

and an average intron length of 12,522 bp, values which are comparable to those of the published migratory locust genome (Wang *et al.*, 2014). Although both locust genomes have longer pre-mRNAs with bigger introns and more exons than the *Drosophila melanogaster* genome (Adams *et al.*, 2000), their average CDS and exon length are in fact shorter (Figure 2 and Table 3). The BUSCO assessment of the current gene set (protein mode) shows that it includes 79.4% complete genes in the insecta_odb10 dataset (Simão *et al.*, 2015), which closely matches the result from the BUSCO genome completeness assessment (genome mode) of 80.9% (Extended data, Supplementary Table S5 (Verlinden *et al.*, 2020)). The BUSCO assessment of the predicted genes in the desert locust genome shows fewer complete genes than for the published *Locusta migratoria* and *Drosophila melanogaster* genomes (Figure 2). Among the 18,815 predicted genes in the desert locust genome, 13,646 (72.53%) obtained at least one functional assignment based on similarity to known proteins in the databases. Pfam domain information could be added to 10,395 (55.25%) predicted genes, and 6,470 (34.39%) predicted genes could be assigned a GO term (Extended data, Supplementary Table S6 (Verlinden *et al.*, 2020)).

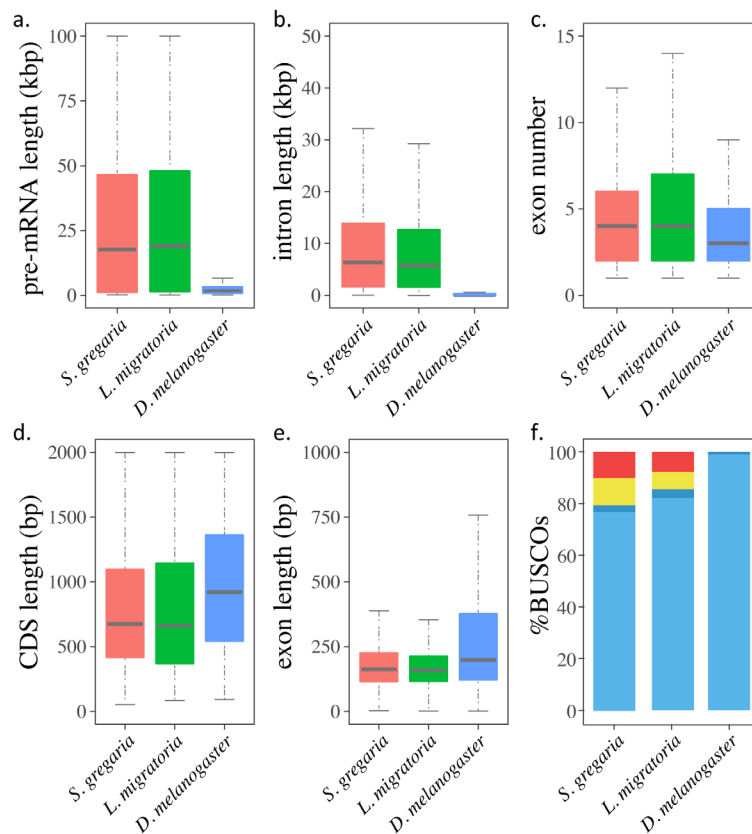


Figure 2. Gene characteristics and BUSCO assessment in the genomes of the desert locust, *Schistocerca gregaria*, the migratory locust, *Locusta migratoria* (Wang *et al.*, 2014) and the fruit fly, *Drosophila melanogaster* (Adams *et al.*, 2000). (a-e) Boxplots of (a) pre-mRNA lengths; (b) intron lengths; (c) exon numbers; (d) coding sequence (CDS) lengths; and (e) exon lengths in the three genomes. (f) BUSCO assessments of the gene sets in the three genomes. The stacked bars indicate the percentages of genes that are complete (light blue), duplicated (dark blue), fragmental (yellow) and missed (red).

Table 3. Summary statistics on gene information for the desert locust, *Schistocerca gregaria*, and the migratory locust, *Locusta migratoria* (Wang *et al.*, 2014).

	<i>Schistocerca gregaria</i>	<i>Locusta migratoria</i>
Genome		
Size (bp)	8,817,834,205	6,524,990,357
Scaffold N50 (bp)	157,705	322,700
GC content	0.406	0.407
Gene		
Total gene number	18,815	17,307
Average pre-mRNA Length (bp)	54,426	54,341
Average CDS length (bp)	1,137	1,160
Average intron length (bp)	12,522	11,159
Average exon length (bp)	216	201
Average exon number per gene	5.26	5.77

Scaffold N50, the sequence length of the shortest scaffold at 50% of the total genome length;

Conclusions

Here, we present the first draft genome sequence of the desert locust, *Schistocerca gregaria*, a swarming pest species with significant socio-economic and ecological impact. With the current locust crisis in mind, it should be clear that despite ongoing monitoring and control operations, we are still in urgent need of more locust research to foster development of effective management strategies. Sequencing and assembling the desert locust genome has been both challenging and groundbreaking due to the enormous size of the genome and its extremely large proportion of repetitive elements. The desert locust genome is the largest insect genome sequenced and assembled to date. As is the case for the second and third largest assembled insect genomes, the expanded genome size is caused by accumulation of repetitive regions and intron elongation (*Locusta migratoria*, 6.5 Gb; Wang *et al.*, 2014; *Clitarchus hookeri*, 4.2 Gb; Wu *et al.*, 2017). Sequencing the desert locust genome is an important step to advance our knowledge of these animals. It will enable future studies to examine the very complex relationship between environmental cues and phenotypic plasticity, and in particular the question of how this is regulated at the molecular level. A better understanding of the desert locust's molecular biology will facilitate the development of novel, more sustainable strategies for controlling these pests.

Data availability

Underlying data

European Nucleotide Archive: First draft genome of *Schistocerca gregaria*, a swarm forming grasshopper species.

Accession number PRJEB38779; <https://identifiers.org/ena.embl:PRJEB38779>.

This accession contains all genome and transcriptome data. The annotations are also available via the ORCAE platform (<https://bioinformatics.psb.ugent.be/orcae/overview/Schgr>).

Extended data

Figshare: First draft genome assembly of the desert locust, *Schistocerca gregaria* - extended data. <https://doi.org/10.6084/m9.figshare.12654026.v1> (Verlinden *et al.*, 2020).

This project contains the following extended data:

- Supplementary Methods (DOCX). Containing details of Animal material, Genomic DNA extraction, Library construction, sequencing for RNA-Seq and *de novo* transcriptome assembly.
- Supplementary Table S1 (DOCX). Available Polyneopteran genomes (incl. *Schistocerca gregaria* for comparison).
- Supplementary Table S2 (DOCX). Software parameter settings.
- Supplementary Table S3 (DOCX). Transfer RNA (tRNA), microRNA (miRNA), small nuclear RNA (snRNA) and ribosomal RNA (rRNA) content of the desert locust genome.
- Supplementary Table S4 (DOCX). Desert locust genome annotation details.

- Supplementary Table S5 (DOCX). BUSCO assessments for the genomes of the desert locust, *Schistocerca gregaria*, and the migratory locust, *Locusta migratoria* (Wang *et al.*, 2014).
- Supplementary Table S6 (DOCX). Functional annotation of the proteome of the desert locust.

Extended data are available under the terms of the [Creative Commons Attribution 4.0 International license](#) (CC-BY 4.0).

Acknowledgements

Evelien Herinckx (KU Leuven) for technical support in desert locust rearing; Evert Bruyninckx (KU Leuven) for optimizing genomic DNA extraction. Ellen De Meester and Sarah De Keulenaer from NxtGNT Belgium for their practical expertise and assistance in the Illumina sequencing experiments. Wim Meert and Stephanie Deman (Genomics core Leuven) for optimizing the PacBio sequencing.

References

- Adams MD, Celniker SE, Holt RA, *et al.*: **The genome sequence of *Drosophila melanogaster***. *Science*. 2000; **287**(5461): 2185–2195.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Aitschul SF, Gish W, Miller W, *et al.*: **Basic local alignment search tool**. *J Mol Biol*. 1990; **215**(3): 403–410.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Badisco L, Huybrechts J, Simonet G, *et al.*: **Transcriptome analysis of the desert locust central nervous system: production and annotation of a *Schistocerca gregaria* EST database**. *PLoS One*. 2011; **6**(3): e17274.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Bryant DM, Johnson K, DiTommaso T, *et al.*: **A tissue-mapped *Axolotl* *de novo* transcriptome enables identification of limb regeneration factors**. *Cell Rep*. 2017; **18**(3): 762–776.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Camacho C, Coulouris G, Avagyan V, *et al.*: **BLAST+: architecture and applications**. *BMC Bioinformatics*. 2008; **10**: 421.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Camacho JPM, Ruiz-Ruano FJ, Martín-Blázquez R, *et al.*: **A step to the gigantic genome of the desert locust: chromosome sizes and repeated DNAs**. *Chromosoma*. 2015; **124**(2): 263–75.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Courcoux G: **Invasions of locusts: a lasting impact**. *Scientific news of the Institut de Recherche pour le Développement*. 2012; **411**.
[Reference Source](#)
- Cullen DA, Cease AJ, Latchininsky AV, *et al.*: **From molecules to management: Mechanisms and consequences of locust phase polyphenism**. *Adv Insect Physiol*. 2017; **53**: 167–285.
[Publisher Full Text](#)
- De Vrejer P, Guilbert N, Mespel-Somps S: **The 1987-89 locust plague in Mali: Evidences of the heterogeneous impact of income shocks on education outcomes**. *No DT/2012/05, Working Papers, DIAL (Développement, Institutions et Mondialisation)*. 2012; 48p.
[Reference Source](#)
- Fox DP: **A non-doubling DNA series in somatic tissues of the locusts *Schistocerca gregaria* (Forskål) and *Locusta migratoria* (Linn.)**. *Chromosoma*. 1970; **29**(4): 446–461.
[PubMed Abstract](#)
- Grabherr MG, Haas BJ, Yassour M, *et al.*: **Full-length transcriptome assembly from RNA-seq data without a reference genome**. *Nat Biotechnol*. 2011; **29**(7): 644–652.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Griffiths-Jones S, Bateman A, Marshall M, *et al.*: **Rfam: an RNA family database**. *Nucleic Acids Res*. 2003; **31**(1): 439–441.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Haas BJ, Delcher AL, Mount SM, *et al.*: **Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies**. *Nucleic Acids Res*. 2003; **31**(19): 5654–5666.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Haas BJ, Papanicolaou A, Yassour M, *et al.*: **De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis**. *Nat Protoc*. 2013; **8**(8): 1494.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Haas BJ, Salzberg SL, Zhu W, *et al.*: **Automated eukaryotic gene structure annotation using EvidenceModeler and the Program to Assemble Spliced Alignments**. *Genome Biol*. 2008; **9**(1): R7.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Hoff KJ, Lomsadze A, Borodovsky M, *et al.*: **Whole-genome annotation with BRAKER**. *Methods Mol Biol*. 2019; **1962**: 65–95.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- John B, Hewitt, GM: **Karyotype stability and DNA variability in the Acrididae**. *Chromosoma*. 1966; **20**: 155–172.
[Publisher Full Text](#)
- Kim D, Langmead B, Salzberg SL: **HISAT: a fast spliced aligner with low memory requirements**. *Nat Methods*. 2015; **12**(4): 357–360.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Koren S, Walenz BP, Berlin K, *et al.*: **Canu: scalable and accurate long-read assembly via k-mer weighting and repeat separation**. *Genome Res*. 2017; **27**(5): 722–736.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Korf I: **Gene finding in novel genomes**. *BMC Bioinformatics*. 2004; **5**(1): 59.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Lorenz R, Bernhart SH, Höner Zu Siederdisen C, *et al.*: **ViennaRNA Package 2.0**. *Algorithms Mol Biol*. 2011; **6**(1): 26.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence**. *Nucleic Acids Res*. 1997; **25**(5): 955–964.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Magrane M: **UniProt Knowledgebase: a hub of integrated protein data**. *Database (Oxford)*. 2011; bar009.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Majoros WH, Pertea M, Salzberg SL: **TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders**. *Bioinformatics*. 2004; **20**(16): 2878–2879.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Meynard CN, Lecoq M, Chapuis MP, *et al.*: **On the relative role of climate change and management in the current desert locust outbreak in East Africa**. *Glob Chang Biol*. 2020; **26**(7): 3753–3755.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Misof B, Liu S, Meusemann K, *et al.*: **Phylogenomics resolves the timing and pattern of insect evolution**. *Science*. 2014; **346**(6210): 763–767.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Mitchell AL, Attwood TK, Babbitt PC, *et al.*: **InterPro in 2019: improving coverage, classification and access to protein sequence annotations**. *Nucleic Acids Res*. 2019; **47**(D1): D351–D360.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Nawrocki EP, Eddy SR: **Infernal 1.1: 100-fold faster RNA homology searches**. *Bioinformatics*. 2013; **29**(22): 2933–2935.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- O’Connell J, Schulz-Trieglaff O, Carlson E, *et al.*: **NxTrim: Optimized trimming of Illumina mate pair reads**. *Bioinformatics*. 2015; **31**(12): 2035–2037.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Pener MP, Simpson SJ: **Locust phase polyphenism: an update**. *Adv Insect Physiol*. 2009; **36**: 1–272.
[Publisher Full Text](#)
- Pertea M, Pertea GM, Antonescu CM, *et al.*: **StringTie enables improved reconstruction of a transcriptome from RNA-seq reads**. *Nat Biotechnol*. 2015; **33**(3): 290–5.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Schoofs L, Veelaert D, Vanden Broeck J, *et al.*: **Peptides in the locusts, *Locusta migratoria*, and *Schistocerca gregaria***. *Peptides*. 1997; **18**(1): 145–56.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Simão FA, Waterhouse RM, Ioannidis P, *et al.*: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs**. *Bioinformatics*. 2015; **31**(19): 3210–3212.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Simpson JT, Wong K, Jackman SD, *et al.*: **ABYSS: a parallel assembler for short read sequence data**. *Genome Res*. 2009; **19**(6): 1117–1123.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Slater GSC, Birney E: **Automated generation of heuristics for biological**

sequence comparison. *BMC Bioinformatics*. 2005; 6 (1): 31.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Stanke, M, Keller, O, Gunduz, I, *et al.*: **AUGUSTUS: *ab initio* prediction of alternative transcripts.** *Nucleic Acids Res.* 2006; 34(Web Server issue): W435–W439.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Symmons PM, Cressman K: **Desert Locust Guidelines.** Second edition. Food and Agriculture Organization of the United Nations (Rome). 2001.

[Reference Source](#)

Tarailo-Graovac M, Chen N: **Using RepeatMasker to identify repetitive elements in genomic sequences.** *Curr Protoc Bioinformatics*. 2009; Chapter 4: Unit 4.10.

[PubMed Abstract](#) | [Publisher Full Text](#)

Tav C, Tempel S, Poligny L, *et al.*: **miRNAFold: a web server for fast miRNA precursor prediction in genomes.** *Nucleic Acids Res.* 2016; 44(W1): W181–W184.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

The UniProt Consortium: **UniProt: a worldwide hub of protein knowledge.** *Nucleic Acids Res.* 2019; 47(D1): D506–D515.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

van Huis A, Van Itterbeeck J, Klunder H, *et al.*: **Edible insects: future prospects for food and feed security.** *FAO Forestry Paper*. 2013; 171: 187.

[Reference Source](#)

Verlinden H, Badisco L, Marchal E, *et al.*: **Endocrinology of reproduction and phase transition in locusts.** *Gen Comp Endocrinol.* 2009; 162(1): 79–92.

[PubMed Abstract](#) | [Publisher Full Text](#)

Verlinden H, Sterck L, Li J, *et al.*: **First draft genome assembly of the desert**

locust, *Schistocerca gregaria* - extended data. *figshare.* Journal contribution. 2020. <http://www.doi.org/10.6084/m9.figshare.12654026.v1>

Wang X, Fang X, Yang P, *et al.*: **The locust genome provides insight into swarm formation and long-distance flight.** *Nat Commun.* 2014; 5: 2957.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Wang Y, Jiang F, Wang H, *et al.*: **Evidence for the expression of abundant microRNAs in the locust genome.** *Sci Rep.* 2015; 5: 13608.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Warren RL, Yang C, Vandervalk BP, *et al.*: **LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads.** *Gigascience.* 2015; 4(1): 35.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Wilmore PJ, Brown AK: **Molecular properties of orthopteran DNA.** *Chromosoma.* 1975; 51(4): 337–345.

[PubMed Abstract](#) | [Publisher Full Text](#)

Wipfler B, Letsch H, Frandsen PB, *et al.*: **Evolutionary history of Polyneoptera and its implications for our understanding of early winged insects.** *Proc Natl Acad Sci USA.* 2019; 116(8): 3024–3029.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Wu C, Twort VG, Crowhurst RN, *et al.*: **Assembling large genomes: analysis of the stick insect (*Clitarchus hookeri*) genome reveals a high repeat content and sex-biased genes associated with reproduction.** *BMC Genomics.* 2017; 18(1): 884.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Xu Z, Wang H: **LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons.** *Nucleic Acids Res.* 2007; 35(Web Server issue): W265–W268.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research