Spodoptera littoralis genome mining brings insights on the dynamic of expansion of gustatory receptors in polyphagous noctuidae

To cite this version:
Camille Meslin, Pauline Mainet, Nicolas Montagné, Stéphanie Robin, Fabrice Legeai, et al.. Spodoptera littoralis genome mining brings insights on the dynamic of expansion of gustatory receptors in polyphagous noctuidae. G3, 2022, 12 (8), pp.jkac131. 10.1093/g3journal/jkac131. hal-03713321

HAL Id: hal-03713321
https://hal.inrae.fr/hal-03713321
Submitted on 4 Jul 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License
**Spodoptera littoralis** genome mining brings insights on the dynamic of expansion of gustatory receptors in polyphagous noctuidae

Camille Meslin, Pauline Mainet, Nicolas Montagné, Stéphanie Robin, Fabrice Legeai, Anthony Bretauudeux, J. Spencer Johnston, Fotini Koutroumpa, Emma Persyn, Christelle Monsempé, Marie-Christine François, Emmanuelle Jacquin-Joly

1 INRAE, Sorbonne Université, CNRS, IRD, UPEC, Université de Paris, Institut d’Ecologie et des Sciences de l’Environnement de Paris (EES-Paris), 78026 Versailles, France,
2 INRAE, UMR Institut de Génétique, Environnement et Protection des Plantes (IGEPP), Bioinformatics Platform for Agroecosystems Anthropods (BIPAA), Campus Beaulieu, 35042 Rennes, France,
3 INRA, IRISA, GenOuest Core Facility, Campus de Beaulieu, Rennes 5042, France,
4 Department of Entomology, Texas A&M University, College Station, TX 77843, USA,
5 CIRAD, UMR PVBMT, Réunion, France,
6 Present address: INRAE, Université Tours, Infectiologie et Santé Publique (ISP), 37380 Nouzilly, France

*Corresponding author: INRAE, Sorbonne Université, CNRS, IRD, UPEC, Université de Paris, Institut d’Ecologie et des Sciences de l’Environnement de Paris (EES-Paris), 78026 Versailles, France. Email: emmanuelle.joly@inrae.fr

**Abstract**

The bitter taste, triggered via gustatory receptors, serves as an important natural defense against the ingestion of poisonous foods in animals, and the increased host breadth is usually linked to an increase in the number of gustatory receptor genes. This has been especially observed in polyphagous insect species, such as noctuid species from the Spodoptera genus. However, the dynamic and physical mechanisms leading to these gene expansions and the evolutionary pressures behind them remain elusive. Among major drivers of genome dynamics are the transposable elements but, surprisingly, their potential role in insect gustatory receptor expansion has not been considered yet. In this work, we hypothesized that transposable elements and possibly positive selection would be involved in the highly dynamic evolution of gustatory receptor in Spodoptera spp. We first sequenced de novo the full 465 Mb genome of *S. littoralis*, and manually annotated the main chemosensory genes, including a large repertoire of 373 gustatory receptor genes (including 19 pseudogenes). We also improved the completeness of *S. frugiperda* and *S. litura* gustatory receptor gene repertoires. Then, we annotated transposable elements and revealed that a particular category of class I retrotransposons, the SINE transposons, was significantly enriched in the vicinity of gustatory receptor gene clusters, suggesting a transposon-mediated mechanism for the formation of these clusters. Selection pressure analyses indicated that positive selection within the gustatory receptor gene family is cryptic, only 7 receptors being identified as positively selected. Altogether, our data provide a new good quality Spodoptera genome, pinpoint interesting gustatory receptor candidates for further functional studies and bring valuable genomic information on the mechanisms of gustatory receptor expansions in polyphagous insect species.

**Keywords:** *Spodoptera littoralis*; gustatory receptors; transposable elements

**Introduction**

Animals rely heavily on their sense of taste to discriminate between harmful poisonous foods, usually through the detection of bitter compounds, and beneficial sustenance. Interestingly, narrowness of food diets in animals is usually linked to a decreased number of gustatory receptors (GRs), in both mammals such as the blood-feeder bats (Hong and Zhao 2014), and in insects such as the body louse (Kirkness et al. 2010)—an obligate ectoparasite of human—the fig wasp Ceratosolen solmsi (Xiao et al. 2013)—specialized on Ficus—and many Lepidoptera specialist feeders, although mammals and insect GRs are unrelated. Reversely, the increased host breadth is usually linked to GR expansions. This has been especially observed in polyphagous insects, including omnivorous species such as the American cockroach *Periplaneta americana* (Li et al. 2018) and herbivorous species such as noctuid species (Xu et al. 2016, Cheng et al. 2017, Gouin et al. 2017).

In polyphagous noctuids, the sequencing of the genomes of *Spodoptera frugiperda* and *Spodoptera litura* revealed GR repertoires of 231 and 237 genes (Cheng et al. 2017; Gouin et al. 2017), respectively, more than twice as much compared with other monophagous and oligophagous Lepidoptera species (Bombyx mori: 69 genes, Heliconius melpomene: 73 genes) (The International Silkworm Genome Consortium 2008; The Heliconius Genome Consortium 2012; Briscoe et al. 2013; Guo et al. 2017), suggesting that the number of GRs has greatly increased during evolution in polyphagous
Lepidoptera via gene tandem duplication. The genomic architecture of the GR family is thus well known in these species and, together with previous studies, it supports the evidence that the family evolved under a birth-and-death model as well as under different selective pressures depending on the clade considered (Nei and Rooney 2005; Briscoe et al. 2013; Almeida et al. 2014; Suzuki et al. 2018). Most of these GRs belong to clades grouping the so-called “bitter” receptors, but in fact the function of the majority of these GRs remains enigmatic. Although the bitter GR class exhibits the most dynamic evolution, the mechanisms leading to GR expansions and the evolutionary pressures behind them remain elusive. Among major drivers of genome dynamics are the transposable elements (TEs). TEs are very diverse and are distributed along genomes in a nonrandom way. Similar or identical TEs can induce chromosomal rearrangements such as deletions, insertions, and even duplications, features that are frequent in multicogene families such as GRs. Surprisingly, their potential role in insect GR expansion has not been considered yet.

In order to study GR evolution and the potential role of TEs in GR expansion in more detail, we sequenced an additional genome of a Spodoptera species: Spodoptera littoralis. So far, only 38 GRs were identified (Poivet et al. 2013; Walker et al. 2019; Koutroumpa et al. 2021) in S. littoralis whereas several hundreds of GRs were annotated in its counterparts S. litura and S. frugiperda. To investigate this singularity, we report here the sequencing of the S. littoralis genome, its full assembly, functional automatic annotation and expert annotation of the main chemosensory gene families, namely soluble carrier proteins [odorant-binding proteins (OBPs) and chemosensory proteins (CSPs)] (Pelosi et al. 2018) and the 3 major families of insect chemosensory receptors [odorant receptors (ORs); ionotopic receptors (IRs), and GRs] (Robertson 2019). With a particular focus on gustation, we also reannotated GRs in S. litura and S. frugiperda. Then, we analyzed the evolutionary history of GRs, by looking at the enrichment for TEs in the vicinity of GRs and by analyzing selective pressures acting on the different GR clades.

### Materials and methods

#### Estimation of S. littoralis genome size

The genome size of S. littoralis was estimated using flow cytometry. Genomic size estimates were produced as described before (Johnston et al. 2019). In brief, the head of a S. littoralis adult male along with the head of a female Drosophila virilis standard (1C = 328 Mbp; 1C = amount of DNA in the genome of homogamic sex) (Greilhuber et al. 2005) were placed into 1 ml of Galbraith buffer in a 2-ml Kontes Dounce and ground with 15 strokes of the A pestle. The released nuclei were filtered through a 40-µm nylon filter and stained with 25 µg/ml propidium iodide for 2 h in the dark at 4°C. The average red fluorescence of the 2C nuclei was scored with a Partec C flow cytometer emitting at 514 nm. The 1C genome size of S. littoralis was estimated as (average red fluorescence of the 2C S. littoralis peak)/(average fluorescence of the 2C D. virilis peak) × 328 Mbp.

#### Spodoptera littoralis genome sequencing and assembly

##### Biological material and genomic DNA extraction

Whole genomic DNA was extracted from 2 male larvae obtained after 2 generations resulting from a single pair of S. littoralis originating from an inbred laboratory colony maintained in INRAE Versailles since 2000s on a semiartificial diet (Poitout and Bues 1974) at 24 ± 2°C and 65 ± 5% relative humidity under 16:8 h light-dark photoperiod. The sex of individuals was verified by checking for presence of testis. The gut was removed and DNA extraction was performed from whole, late-stage larvae using Qiagen Genomic-tip 500/G (Qiagen Inc., Chatsworth, CA, USA). A total of 30 µg of genomic DNA were obtained.

#### Sequencing

Different types of libraries were generated for 2 sequencing technologies: Illumina and PacBio. For Illumina sequencing, 5 libraries were prepared and constructed according to the Illumina manufacturer’s protocol (1 library of 170, 1 of 250, and 3 of 500 bp). Illumina sequencing was performed at the BGI-tech facilities (Shenzhen, China) on a HiSeq2500 machine. Around 68 Gb were obtained, representing 144× of the estimated genome size (470 Mb) (Supplementary Data 1). The raw reads were filtered at BGI to remove adapter sequences, contaminations, and low-quality reads and the quality of all raw reads was assessed using FASTQC (Andrews S, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). PacBio sequencing was performed at GenoScreen (Lille, France) by the SMRT sequencing technology on 9 SMRTcell RSII, generating 2 846 820 reads. Around 16 Gb were obtained, representing 34× of the estimated genome size (Supplementary Data 1). High-quality sequences were obtained by generating circular consensus sequencing.

##### Genome assembly

A first assembly was done using Platanus (v1.2.1) (Kajitani et al. 2014) with Illumina data. A second assembly was obtained by doing scaffolding with SSPACE-LR (modified) (Boetzer and Pirovano 2014) using PacBio data and gap filling using GapCloser (Boetzer and Pirovano 2012). These second assembly was evaluated using Benchmarking Universal Single-Copy Orthologue (BUSCO v5.2.2) (Simão et al. 2015) with a reference set of 5,286 genes conserved in Lepidoptera.

##### Structural and functional genome annotation

Structural automatic genome annotation was done with BRAKER (v1.11) (Hoff et al. 2016) using all RNAseq data described in Supplementary Data 1. RNAseq libraries were sequenced from different larvae and adult tissues from males and females including the proboscis, palps, legs, and ovipositor and sequenced by Illumina (Supplementary Data 1) (Poivet et al. 2013; Walker et al. 2019; Koutroumpa et al. 2021). Reads were trimmed using Trimomatic (v0.36) (Bolger et al. 2014) with the following parameters: ILLUMINACLIP: TruSeq2-PE.fa: 2:30:10, LEADING: 3, TRAILING: 3, SLIDINGWINDOW: 4:15, MINLEN: 36. Trimmed reads were mapped on the genome assembly using STAR (v5.2a) (Dobin et al. 2013) with the default parameters except for the following parameters: outFilterMultimapNmax = 5, outFilterMismatchNmax = 3, alignIntronMin = 10, alignIntronMax = 50,000, and alignMates GapMax = 50,000. As done for the genome assembly, gene annotation was evaluated using BUSCO v5.2.2 with a reference set of 5,286 proteins conserved in Lepidoptera. Putative functions of predicted proteins were assigned using blastp (v2.6.0) against GenBank NR (nonredundant GenBank CDS translations + PDB + SwissProt + PIR + PRF) release 2017 September, and interproscan v5.13-52.0 against Interpro. Associated GO terms were collected from blast NR and interproscan results with blast2GO (v2.5).

##### Annotation of OBPs, CSPs, ORs, and IRs

The annotation of genes encoding soluble transporters (OBPs and CSPs), ORs, and IRs was performed using known sequences from other species with their genome sequenced (S. frugiperda, S. litura, S. frugiperda, S. litura, and S. littoralis).
Annotation and enrichment analysis of TEs around chemosensory receptor genes in Spodoptera species

The annotation of TEs in S. littoralis genome was performed using REFET (Galaxy Lite v2.5). The TDenovo pipeline (Flutre et al. 2011) was used to identify consensus sequences representative of each type of repetitive elements. Only contigs of a length $\geq 10\text{kb}$ were used as input for the pipeline. Consensus sequences were built only if at least 3 similar copies were detected in the genome. The TEannot pipeline (Quesneville et al. 2005) was then used to annotate all repetitive elements in the genome using the library of TE consensus and to build an NR library in which redundant consensus were eliminated (length $\geq 98\%$, identity $\geq 95\%$). The NR library of TEs was finally used to perform the S. littoralis genome annotation with the TEannot pipeline.

The tool Locus Overlap Analysis (LOLA) within the R package Bioconductor (Sheffield and Bock 2016) was used to test for enrichment of TEs within the genomic regions containing chemosensory receptor genes (ORs and GRs) in both S. littoralis and S. frugiperda. To run LOLA with data from S. littoralis, 3 types of datasets were created. The first dataset, the query set, contained genomic regions of 10\text{kb} around each chemosensory receptor gene. Since these genes were mostly organized in clusters within the genome, regions with overlap were combined leading to the creation of 114 chemosensory regions for the GRs and 63 regions for the ORs. The second dataset, the region universe, contained all genomic regions from the genome. The region universe was created by retrieving the gene coordinates from the Official Gene Set from both genome (OGS3.0 for S. littoralis, OGS2.2 for S. frugiperda), and expanding to 10\text{kb} around each gene. Similarly to the GR regions, regions with overlap were combined. This led to the creation of 14,072 genetic regions for S. littoralis and 11,053 genetic regions for S. frugiperda. The last dataset, the reference dataset, contained the coordinates of TEs previously identified by the REPET analysis. The enrichment in TE content within the chemosensory regions and the control regions were then compared using LOLA using a Fisher’s Exact Test with false discovery rate correction ($q$-value) to assess the significance of overlap in each pairwise comparison. The same method was used using S. frugiperda TEs, previously annotated using the same tool REPET (Gouin et al. 2017), as well as chemosensory receptor reannotations from the present work and led to the creation of 191 chemosensory regions for the GRs and 88 regions for the ORs.

Evolutionary analyses

Phylogenetic tree reconstructions

Chemosensory-related protein trees were constructed for OBPs, CSPs, ORs, IRs, and GRs. For GRs, the phylogeny was built using GR amino acid sequences from different Lepidoptera species with various diets. In order to take into account the whole repertoire of GRs in our analysis, only species in which the GRs were annotated following whole-genome sequencing were considered. The dataset contained GRs from polyphagous (S. littoralis, S. litura, and S. frugiperda), oligophagous (H. melpomene—73 GRs, Manduca sexta—45 GRs) and monophagous species (B. mori—72 GRs). The multiple sequence alignment of all GR amino acid sequences (except short partial sequences) was performed with ClustalO (Sievers et al. 2011) and the phylogeny was reconstructed using PhyML 3.0 (Guindon and Gascuel 2003) (http://www.atgc-montpellier.fr/phyml/) with the automatic selection of the best substitution model by SMS (Lefort et al. 2011) and the phylogeny was reconstructed using PhyML 3.0 (Guindon and Gascuel 2003) (http://www.atgc-montpellier.fr/phyml/) with the automatic selection of the best substitution model by SMS (Lefort et al. 2011). The CO2 and sugar receptor clades were used to root the tree. The resulting phylogenetic tree was edited using FigTree v1.4.2 (https://github.com/rambaut/figtree)
tree) and Inkscape 0.92 (https://inkscape.org/fr/). Branch supports were estimated using the approximate likelihood-ratio test (Anisimova and Gascuel 2006) implemented at http://www.atgc-montpellier.fr/phylm/. For other gene families, sequences from various Lepidoptera species were retrieved and aligned with S. littoralis sequences using MAFFT (Katoh and Standley 2013). The reconstruction of the phylogenetic trees was carried out the same way as for the GRs, and the OR tree was rooted with the Orco clade.

Tree reconciliation
Estimates of gains and losses of GR genes across the Noctuidae were inferred using the reconciliation methods implemented in Notung v2.6 (Stolzer et al. 2012; Darby et al. 2016). The species tree was generated using TimeTree.org (Kumar et al. 2017) and the gene tree was the reconstructed phylogeny of the GRs generated by PhyML.

Evolutionary pressures
The codeml software of the package PAML was used to infer selective pressures (Yang 2007). Because of the high divergence between GRs across the phylogeny, selective pressures were inferred on 13 subtrees extracted from the GR phylogeny in order to minimize the ratio of synonymous substitutions. For each subtree, an alignment of the protein sequences was performed using MAFFT, converted to codon alignment using PAL2NAL (Suyama et al. 2006), and a phylogenetic tree was reconstructed based on the protein sequence alignment. Sequences introducing large gaps in the alignment were removed in order to compute codeml on the largest alignment possible. To estimate the selective pressures acting on the evolution of the lepidopteran GR genes, the “m0 model” from codeml of the PAML package was computed on the 13 subtrees to estimate the global ω (ratio of nonsynonymous substitutions dN/ratio of synonymous substitutions dS) (Yang et al. 2000). The ω value reflects the mode of evolution, with ω > 1 indicating positive selection, ω < 1 indicating purifying selection, and ω = 1 indicating neutral evolution. To further infer positive selection, 2 comparisons between evolutionary models were conducted. First, the comparison between M8 and M8a models can detect positive selection acting on sites, i.e. columns of the alignment (Swanson et al. 2005; Wong et al. 2004). This comparison was conducted only when the global ω calculated from the m0 model was >0.3. The second comparison between branch-site model A and its neutral counterpart can detect positive selection acting on particular sites on a specific lineage (Zhang et al. 2005), a method reported to be more powerful than other comparisons implemented in PAML to detect episodic positive selection (Yang and Dos Reis 2011). Here, we tested all the terminal branches of the trees for which both the global ω was elevated and the comparison between models M8 and M8a statistically significant. Since many branches were tested for each tree, a correction for multiple testing to control for false discovery rate was applied: the q-value (q-value R package version 2.22.0; Storey et al. 2013). In the case of a statistically significant q-value (<0.05), positively selected sites were inspected for possible artifacts due to partial sequences or misalignment.

Putative functional assignment
In order to assign putative functions to several candidate SlitGRs, both their phylogenetic position and theoretical 3D structure were analyzed. For the theoretical structures, the AlphaFold algorithm (Jumper et al. 2021) was used to model candidate SlitGRs as well as their B. mori ortholog GRs with known function: BmorGR9 and BmorGR66. Structures were then compared between orthologs using the MatchMaker tool of Chimera and the root mean square deviation (RMSD) computed using the same tool (Pettersen et al. 2004). Docking of D-fructose was performed on both BmorGR9 and SlitGR9 using the Webina webserver (https://durrantlab.pitt.edu/webina/) (Kochnev et al. 2020).

Results and discussion
Genome assembly and automatic annotation of the S. littoralis genome
The first assembly of S. littoralis (v1.0), obtained with short Illumina reads, contained 123,499 scaffolds with a N50 of 18 kb and an assembly total size around 470 Mb. The second assembly (v2.0), obtained with a combination of both short Illumina and long PacBio reads, contained 28,891 scaffolds, with a N50 of 64 kb and an assembly total size around 465 Mb (Table 1). The genome size of S. littoralis was in good correlation with flow cytometry evaluation (470 Mb). The BUSCO analysis revealed that the second assembly contained more than 97% of complete BUSCO genes, with more than 95% of them being present in single-copy (Table 2). This second assembly was then used as the final assembly in all the following analyses. A total of 35,801 genes were predicted using BRAKER (OGS3.0_20171108). The number of genes annotated in lepidopteran genomes is usually much lower (around 15,000 genes in general). This unusual number is probably due to duplicates, as revealed by the BUSCO analysis (10% duplicated annotated genes). However, almost 97% of BUSCO proteins were complete, with more than 86% being present in single-copy (Table 2). These data show that the S. littoralis genome assembly is of good quality, thus allowing for accurate comparison with other Spodoptera genomes.

OBP, CSP, OR, and IR chemosensory gene repertoires were of comparable size among Spodoptera spp.
To have a full view of the S. littoralis chemosensory gene repertoires, we manually curated all the major chemosensory-related gene families, including soluble carrier proteins (OBPs and CSPs), proposed to facilitate chemical transfer to chemosensory receptors (Pelosi et al. 2018), and the membrane bound receptors (ORs: 7 transmembrane receptors expressed in the membrane of olfactory sensory neurons; GRs: 7 transmembrane receptors hosted by taste neurons; and IRs: 3 transmembrane proteins sensing acids and amines).

The genome of S. littoralis contained 23 CSP genes, all of them encoding full-length sequences with a signal peptide. This number of genes is similar to the 22 CSP genes annotated in S. frugiperda (Gouin et al. 2017) and the 23 CSP genes annotated in S. litura (Cheng et al. 2017). Among all these sequences, 16 CSP genes are 1:1 orthologs between the 3 Spodoptera species included in the tree while 11 CSP genes are 1:1 orthologs with BmorCSPs. The pairwise alignment of the CSP sequences was done with MAFFT (Katoh and Standley 2013) and the alignment was used with the MatchMaker tool of Chimera and the root mean square deviation (RMSD) computed using the same tool (Pettersen et al. 2004). Docking of D-fructose was performed on both BmorGR9 and SlitGR9 using the Webina webserver (https://durrantlab.pitt.edu/webina/) (Kochnev et al. 2020).

Table 1. Statistics of the S. littoralis genome assemblies.

<table>
<thead>
<tr>
<th></th>
<th>Slit genome v1.0</th>
<th>Slit genome v2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of scaffolds</td>
<td>123,499</td>
<td>28,891</td>
</tr>
<tr>
<td>Total size of scaffolds</td>
<td>470 Mb</td>
<td>465 Mb</td>
</tr>
<tr>
<td>Longest scaffold</td>
<td>236 kb</td>
<td>816 kb</td>
</tr>
<tr>
<td>N50 scaffold length</td>
<td>18 kb</td>
<td>64 kb</td>
</tr>
<tr>
<td>scaffold %N</td>
<td>0.41</td>
<td>0.92</td>
</tr>
</tbody>
</table>
missing GRs in the genome. In view of these data, we searched for possible process, we annotated an even larger repertoire of GRs in the S. littoralis genome scattered on 110 scaffolds in the genome of S. littoralis (Table 3). We annotated a total of 376 GR genes among 61 scaffolds (Supplementary Data 2), including the obligate IRs and 2 lineage expansions were observed (IR7d and IR25a) (Supplementary Fig. 2). This expansion probably arose from recent tandem duplications at the base of this group as most of the genes of the expansion are organized in synteny in the 3 species (Supplementary Fig. 3).

We annotated 44 IR genes in the S. littoralis genome, 43 of which encoding a full-length sequence with various sizes containing 547–948 amino acids (Supplementary Data 3). In addition to the 2 conserved coreceptors IR8a and IR25a (Croset et al. 2010), we identified 18 candidate antennal IRs putatively involved in odorant detection, 23 divergent IRs putatively involved in taste, and 12 ionotrophic glutamate receptors (iGluRs). The total IR number was similar to the 44 IR genes annotated in S. litura (Zhu et al. 2018) and the 43 IR genes annotated in S. frugiperda (Gouin et al. 2017). Among all these sequences, 43 IR genes are 1:1 orthologs between the 3 Spodoptera species (IR100g was missing in S. frugiperda). The phylogenetic tree revealed a clade containing divergent IRs and 2 lineage expansions were observed (IR7d and IR100), likely attributed to gene duplications (Zhu et al. 2018). The number of divergent IRs was much higher in Spodoptera species (S. littoralis: 26, S. litura: 26, S. frugiperda: 25) than in H. melpomene (16) and B. mori (6). By contrast, phylogenetic analysis (Supplementary Fig 4) reveals that S. littoralis antennal IRs retained a single copy within each orthologous group.

We annotated 73 OR genes in the S. littoralis genome scattered among 61 scaffolds (Supplementary Data 2), including the obligatory coreceptor ORco. The number of OR genes in the S. littoralis repertoire was similar to the repertoire of closely related species (69 in S. frugiperda, 73 in S. litura) and other Lepidoptera (64 in D. plexippus, 73 in M. sexta). The phylogenetic tree of ORs is presented in Supplementary Fig. 5.

Altogether, our annotations revealed that OBP, CSP, OR, and IR repertoires were of comparable size among the Spodoptera spp. investigated.

### A highly dynamic evolution of the GR multigene family in Spodoptera species

Newly obtained genomes of polyphagous noctuidae species such as H. armigera (Pearce et al. 2017), S. litura (Cheng et al. 2017), S. frugiperda (Gouin et al. 2017), and A. ipsilon (Wang et al. 2021) revealed an important expansion of GRs in these species, suggesting an adaptation mechanism to polyphagy. Here, using these known GR protein sequences and an iterative annotation process, we annotated an even larger repertoire of GRs in the S. littoralis genome. In view of these data, we searched for possible missing GRs in the S. frugiperda and S. litura genomes to complete their GR repertoires (Table 3). We annotated a total of 376 GR genes scattered on 110 scaffolds in the genome of S. littoralis, and reannotated 417 GRs on 196 scaffolds in the S. frugiperda genome and 293 GRs on 30 scaffolds in S. litura (Supplementary Data 4). When omitting pseudogenes and alleles, the final repertoires of GRs are composed of 325 genes in S. littoralis, 278 GRs in S. frugiperda, and 280 GRs in S. litura. Our GR analysis not only revealed that the full repertoire of S. littoralis GRs is in fact much bigger than previously reported, but also that the GR numbers in S. litura and S. frugiperda have been under evaluated (although the presence of some alleles may over evaluate these numbers). Among these sequences, several were indeed allelic version of previously annotated genes but several new genes were also identified (Table 3). Among these genes, the percentage of complete genes varied between species, from only 41% in S. frugiperda compared with 79% in S. litura while the percentage of complete GRs in S. littoralis was intermediate (73%). The percentage of allelic sequences were also highly variable, probably depending on the heterozygosity level of each considered genome and quality of the assembly (Supplementary Data 5). Indeed, the highest number of alleles was reached in S. frugiperda, a genome with a high level of heterozygosity (Gouin et al. 2017), while alleles were less frequent in the 2 other Spodoptera genomes considered. Multiple partial genes were also part of these repertoires, many of them being located at the boundaries of scaffolds. For the partial genes annotated in the middle of scaffolds, several explanations are possible: either because of misassembly in the region, or because the tools used to annotate failed to identify some exons or even because these partial sequences could be in fact pseudogenes. As previously shown, multiple clusters of ORs were also found in the S. littoralis genome. The 2 main clusters were found on scaffolds 1,414 and 878 that contained each 27 GR genes. The phylogeny reconstructed using the OR sequences from the 3 Spodoptera species as well as those from B. mori (BmorGRs) and H. melpomene (HmelGRs) showed that a few Spodoptera GRs clustered with candidate CO2, sucrose and fructose receptors, while the majority of the Spodoptera GRs were part of the so-called bitter receptor clades. Among the candidate bitter receptor clades, 11 clades were enriched in Spodoptera genes (numbered from A to K in Fig. 1) and encompassed the majority of the 3 Spodoptera GR repertoires (Table 4). When belonging to the same phylogenetic clade, GRs from the same species tend to be located on the same scaffold, supporting the theory of the expansion of these genes by tandem duplications and few gene losses. For the subsequent analysis, only complete and partial genes were considered while pseudogenes were discarded. Four S. littoralis GRs with only 1 exon were identified, clustered on scaffold 67 and belonging to the same phylogenetic clade (Fig. 1). Interestingly, this clade was very conserved with a 1:1 orthology relationship between the 3 Spodoptera species, the SliTuGRs and SfGrGRs being also monoeXonic. All these monoeXonic genes are orthologs with BmorGR53, a single exon gene that is highly expressed at the larval stage but not in the adult (Guo et al. 2017). BmorGR53 is able to detect the bitter taste and feeding deterrent coumarine. It is then likely
that these 4 single exon GRs play an important role in host–plant recognition in *Spodoptera* species as well.

The GR phylogeny served as a basis for the reconciliation of the gene- and species-tree in order to estimate gene gains and losses. The Notung analysis revealed that the ancestral repertoire of GRs of Noctuidae species contained 58 genes (Fig. 2). Given the numbers of GRs annotated in *Spodoptera* species, it is not surprising that the highest gene gains occurred in the ancestor of *Spodoptera* species (296 gene gains). However, even for species with a smaller repertoire of genes such as *B. mori* (70 GRs) and *H. melpomene* (73 GRs), the turnover of genes compared with the ancestors is high (33 and 41 gene gains, 25 and 26 gene losses, respectively).

**Table 3.** GR repertoires of *Spodoptera* species.

<table>
<thead>
<tr>
<th></th>
<th><em>Spodoptera littoralis</em></th>
<th><em>Spodoptera frugiperda</em></th>
<th><em>Spodoptera litura</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of GR previously annotated</td>
<td>38</td>
<td>231</td>
<td>237</td>
</tr>
<tr>
<td>Complete genes</td>
<td>275 (73%)</td>
<td>172 (41%)</td>
<td>231 (79%)</td>
</tr>
<tr>
<td>Partial genes</td>
<td>50 (13%)</td>
<td>106 (25%)</td>
<td>49 (17%)</td>
</tr>
<tr>
<td>Pseudogenes</td>
<td>19 (5%)</td>
<td>22 (5%)</td>
<td>7 (2%)</td>
</tr>
<tr>
<td>Alleles</td>
<td>29 (8%)</td>
<td>117 (28%)</td>
<td>6 (2%)</td>
</tr>
<tr>
<td>Total in this work</td>
<td>373</td>
<td>417</td>
<td>293</td>
</tr>
</tbody>
</table>

*Fig. 1.* Phylogeny of lepidopteran GRs. The dataset included amino acid sequences from *S. littoralis* (Noctuoidea, red), *S. litura* (Noctuoidea, green), *S. frugiperda* (Noctuoidea, orange), *B. mori* (Bombycoidea, blue), and *H. melpomene* (Papilionoidea, cyan). Sequences were aligned using ClustalO and the phylogenetic tree was reconstructed using PHYML. CO₂ receptor candidates as well as sugar receptor candidates are indicated in purple and yellow, respectively. All the other GRs are part of the bitter receptor clades. The star indicates the clade of single-exon GRs. The clade containing putative CO₂ and sugar receptors was used to root the tree. Bootstrap values are indicated for the main clades. The scale bar represents 0.5 amino acid substitutions per site.

*Fig. 2.*
Table 4. Number of Spodoptera putative bitter receptors by expansion clade.

<table>
<thead>
<tr>
<th>Clade</th>
<th>Spodoptera littoralis</th>
<th>Spodoptera frugiperda</th>
<th>Spodoptera litura</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>20</td>
<td>15</td>
<td>14</td>
</tr>
<tr>
<td>B</td>
<td>65</td>
<td>42</td>
<td>44</td>
</tr>
<tr>
<td>C</td>
<td>40</td>
<td>40</td>
<td>33</td>
</tr>
<tr>
<td>D</td>
<td>97</td>
<td>74</td>
<td>89</td>
</tr>
<tr>
<td>E</td>
<td>7</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>F</td>
<td>12</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>G</td>
<td>10</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>H</td>
<td>16</td>
<td>11</td>
<td>16</td>
</tr>
<tr>
<td>I</td>
<td>7</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>J</td>
<td>16</td>
<td>21</td>
<td>18</td>
</tr>
<tr>
<td>K</td>
<td>4</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>294 (90.5%)</td>
<td>238 (86.9%)</td>
<td>251 (89.6%)</td>
</tr>
</tbody>
</table>

The percentages represent the proportion of Spodoptera genes to the total number of GRs annotated in the 3 Spodoptera species (complete + partial genes indicated in Table 3).

Fig. 2. GR gain and loss estimates across lepidopterans. The gene tree of GRs generated using PhyML was reconciled with the species-tree using Notung (Stolzer et al. 2012) to estimate gene gains and losses. Numbers in boxes represent the size of GR repertoire for extant species as well as ancestors at the nodes of the species tree. Gene gains are indicated in red while gene losses are indicated in green. The expansion that occurred in the ancestor of Spodoptera species is indicated in red on a black background.

Annotation of TEs, enrichment analysis, and selection pressure

To get more insights about the mechanisms that led to the formation of massive genomic clusters of GR genes, we looked at (1) whether TEs could be involved and (2) the selective pressures acting on GR genes.

TEs have been shown to be involved in countless mechanisms of evolution in insects, such as insecticide resistance, the evolution of regulatory networks, immunity, climate adaptation (Chen and Li 2007; Rebollo et al. 2012; You et al. 2013; Chuong et al. 2017; Bourque et al. 2018; Singh et al. 2020; Parisot et al. 2021), and some of them have even been domesticated as genes (Maumus et al. 2015). Gene families involved in these traits have been shown to be enriched in TEs and gene family expansions have been correlated with TE content, for instance in termites (Harrison et al. 2018). Interestingly, enrichment in TEs has not been reported for insect GR gene clusters so far. While annotating GRs in the S. littoralis genome, we noticed the frequent co-occurrence of TEs on the same scaffolds. We thus annotated TEs in the S. littoralis genome and calculated their enrichment in the vicinity of GR genes. We also carried out the same enrichment analysis in S. frugiperda genomes, as TE annotation in this last species has been done using the same REPET pipeline as in our study. The de novo constructed library contained 1,089 consensus sequences of TEs and was used to annotate the S. littoralis genome. The repeat coverage for the S. littoralis genome was 30.22%, representing 140 Mb, which is similar to that of S. frugiperda (29.10%), S. litura (31.8%), and S. exigua (33.12%) (Cheng et al. 2017; Gouin et al. 2017; Zhang, Zhang, Yang, et al. 2019). The relative contribution of the different classes of repetitive elements revealed that class I elements were more represented than class II elements (66.96% vs 20.83%), a classical feature of insect genomes (Maumus et al. 2015) (Fig. 3 and Table 5). However, the repartition and proportions between the different classes differed between these species. The class I SINE was the most represented in S. frugiperda (12.52%) (Gouin et al. 2017) while the class I LINE elements were the most represented in both S. litura and S. exigua, although with a lower proportion of all repeated elements (27.73% and 14.81%, respectively). Remarkably, the proportion of LINE elements identified in the S. littoralis genome was the highest reported so far in arthropods (Petersen et al. 2019), accounting for 52.18% of all repetitive elements. In 2 subspecies of the Asian gypsy moth Lymantria dispar, the accumulation of this particular class of TEs was found to be responsible for their large genome size (Hebert et al. 2019), a phenomenon also observed in other insect species (Maumus et al. 2015). The accumulation of the same elements in the S. littoralis genome could explain its larger size compared with its Spodoptera counterparts (465 Mb vs ~400 Mb for S. frugiperda, 438 Mb for S. litura, 408–448 Mb for S. exigua). The second most represented was DNA transposons, class II TIR elements, representing 11.04% of all TEs (Fig. 3 and Table 5).

The enrichment of TEs in the vicinity of GR gene clusters was tested in both S. littoralis and S. frugiperda, and we found that a particular category of class I retrotransposons, a SINE transposon, was significantly enriched in the vicinity of the GRs in the S. littoralis genome (q-value = 0.038), suggesting a transposon-mediated mechanism for the formation of GR clusters (Supplementary Data 6). SINE elements are typically small (80–500 bp) and originate from accidental retrotransposition of various polymerase III transcripts. These elements are nonautonomous, therefore their involvement in the dynamic of the GR multigene family may be related to their potential to induce genome rearrangements via unequal crossing over, hence potential drivers of duplication, as previously shown in other insect species. Given their prevalence in the S. littoralis genome, the potential role of these TEs in the GR family dynamic is probably just one of their numerous functions. The same enrichment analysis performed for the OR loci showed no significant enrichment in both S. littoralis and S. frugiperda (Supplementary Data 6). The involvement of TEs in the expansion of chemosensory gene families is not restricted to S. littoralis and GRs, as similar results were also found in ant genomes for ORs (Schrader et al. 2014; McKenzie and Kronauer 2018), suggesting that TEs are major players in the evolution of insect genomes and in species adaptation (Maumus et al. 2015; Gilbert et al. 2021).

Several studies have shown the importance of positive selection in the evolution of multigene families, especially in chemosensory genes such as ORs and GRs (McBride et al. 2007; Smadja
Positively selected chemoreceptors may be linked to adaptation in Drosophila species (Hickner et al. 2016; Diaz et al. 2018). In the pea aphid, signatures of selection have been identified in chemosensory genes, including GRs and ORs, which may be implicated in the divergence of pea aphid host races. We thus analyzed selective pressures focusing on 13 clades of interest in the Spodoptera GR phylogeny: the potential clade of CO2 receptors, the potential clade of sugar receptors and the 11 expended lineages within the so-called bitter receptor clades. For all 13 clades, we observed low global $\omega$ values ranging from 0.01 to 0.42, with the highest observed for candidate bitter receptor clades. The comparison between models M8 and M8a was statistically significant for clades C, F, and J, indicating a signal of positive selection. Branch-site models on terminal branches of the associated trees were then tested on these clades. For clade J, no GR was revealed as evolving under positive selection. However, for clades C and F, 2 and 5 GRs were identified as positively selected, respectively (Table 6). Within these GR sequences, very few positively selected sites were identified for each gene (between 0 and 3; Supplementary Data 7). This finding is coherent with previous studies showing the same pattern of evolutionary rates (Engsontia et al. 2014; Suzuki et al. 2018), especially in S. frugiperda (Gouin et al. 2017) (3 GRs under positive selection when comparing 2 host strains). Taken together, all positive selection analyses indicate that positive selection within the GR gene family is cryptic and may not play an important role in shaping the evolution of Spodoptera GRs. Anyhow, the few positively selected GRs may be interesting candidates for further functional studies.

**Putative functional assignment of candidate SlitGRs**

The complexity of the evolution of the bitter GRs is reflected by their complex functioning. Indeed, in contrast with the relatively simple OR/Orco association that is the basis for olfaction, the molecular basis for gustation is marked by several characteristics that were recently identified in D. melanogaster. First, some GRs...
have to be coexpressed within the same neuron in order to be able to respond to a stimulus. Second, it seems that GR–GR inhibition can modulate neuron responses. The challenge in the next few years will be to characterize both the response spectra and precise expression patterns of GRs of interest. However, those GRs of interest need to be selected. The present work provides us with some valuable candidates such as the single exon GRs for which the function is known in *B. mori*. Also, it seems that individual GRs can play an important role in the ecology of a species. Among examples are *BmorGR9*, which binds D-fructose without the need of any other GR (Sato et al. 2011; Mang et al. 2016), and *BmorGR66*, whose silencing confers to the monophagous *B. mori* larva the ability to feed from different food sources (Zhang, Zhang, Niu, et al. 2019). We identified their *S. littoralis* orthologs as *SlitGR9* and *SlitGR15*, respectively. We predicted their 3D structures using AlphaFold and compared them with the AlphaFold predicted structures of *B. mori* orthologues. Globally, all the predicted structures obtained reasonable pLDDT scores, especially in the extracellular domain where the putative binding pocket may be located, with higher confidence for *SlitGR9* and *BmorGR9* than for *BmorGR66* and *SlitGR15* (Supplementary Data 8). The RMSD computed between the atoms of the 3D structures of *BmorGR9* and *SlitGR9* ranged from 0.0427 to 26.5 Å (global RMSD: 7.804) (Fig. 4a). While disordered regions, that are difficult to predict, such as the N-terminal end or the loop between transmembrane domains 4 and 5 differ, both structures are strikingly similar, suggesting that *SlitGR9* is likely a D-fructose receptor in *S. littoralis*. Additionally, docking experiments on each structure revealed that their binding pockets are located in the same region of the receptor (Fig. 4b). The ligand of *BmorGR66* is not known, however, this receptor is responsible for the feeding difference of *B. mori* for mulberry leaves (Zhang, Zhang, Niu, et al. 2019). Its ortholog *SlitGR15* is a key candidate for functional studies to test if this GR has an impact on the feeding preference in *S. littoralis* as well. When comparing both full structures, the RMSD was 3.431 Å (Fig. 4c). Interestingly, the main differences between both structures were visible in the extracellular domains of the proteins, suggesting that the binding pockets may differ as well.

Apart from these particular GRs, the neuronal coding of taste via more than 200 genes in species like *Spodoptera* moths is not known. But are all these GRs at play in effective taste sense? In fact, comparison of GR gene repertoires with transcript

---

**Table 6. Selective pressure analysis.**

<table>
<thead>
<tr>
<th>Clade</th>
<th>No. of sequences</th>
<th>$\omega$</th>
<th>M0 ($d_{G}/d_{S}$)</th>
<th>P-value (M8 vs M8a)</th>
<th>Branch-site</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>45</td>
<td>0.34109</td>
<td>NS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>133</td>
<td>0.34146</td>
<td>NS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>86</td>
<td>0.34386</td>
<td>0.044804*</td>
<td>Slit_GR217, Slitu_GR155</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>218</td>
<td>0.29174</td>
<td>/</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>14</td>
<td>0.18639</td>
<td>/</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>33</td>
<td>0.41616</td>
<td>0.000504**</td>
<td></td>
<td>Sfru_GR44, Sfru_GR49, Slit_GR44</td>
</tr>
<tr>
<td>G</td>
<td>28</td>
<td>0.36571</td>
<td>NS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>38</td>
<td>0.31933</td>
<td>NS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>16</td>
<td>0.22375</td>
<td>/</td>
<td></td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>38</td>
<td>0.42257</td>
<td>0.005319**</td>
<td>NS</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>11</td>
<td>0.17393</td>
<td>/</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sugar</td>
<td>27</td>
<td>0.05662</td>
<td>/</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CO₂</td>
<td>11</td>
<td>0.01074</td>
<td>/</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NS, non-significant; /, not calculated.

* p-value <0.05.
** p-value <0.01.

---

**Fig. 4.** Comparison of 3D structures of *BmorGR9* and *BmorGR66* with their respective orthologs, *SlitGR9* and *SlitGR15*. Three-dimensional structures were predicted using AlphaFold2 (Jumper et al. 2021). a, c) *BmorGR9/SlitGR9* and *BmorGR66/SlitGR15* were aligned using matchmaker. The RMSD value computed between both structures is represented here on the *BmorGR9* and *BmorGR66* structures, respectively, in a red (high RMSD) to blue palette (low RMSD). b) Docking of D-fructose in *BmorGR9* (green) and in *SlitGR9* (pink). In (a–c), structures are represented oriented with their extracellular domain at the top and the intracellular domain at the bottom.
Data availability
The assembled genome of S. littoralis as well as the genomic data of Spodoptera litura (v1.0) (Cheng et al. 2017) and S. frugiperda (Corn variant, v3.1) (Gouin et al. 2017) are all publicly available on the BIPAA platform (https://bipaa.genouest.org). In addition, S. littoralis (bioProject PRJNA344815) and S. littoralis genome are available on NCBI (JALJOW0000000000). Files and code for the LOLA analysis are available at dataINRAE (https://data.inrae.fr/dataverse/slitloralis_genome).

Supplemental material is available at G3 online.

Acknowledgments
The S. littoralis genome has been sequenced in the framework of the i5K initiative (http://i5k.github.io/) and the InsectGenomes and innovation program under the Marie Skłodowska-Curie grant agreement no. 764840 for the ITN IGNITE project.

Funding
This work has been funded by INRAE (France), the French National Research Agency (ANR-16-CE21-0002-01 and ANR-16-CE02-0003) and by the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 764840 for the ITN IGNITE project.

Conflicts of interest
None declared.

Literature cited
Chen S, Li X. Transposable elements are enriched within or in close proximity to xenobiotic-metabolizing cytochrome P450 genes. BMC Evol Biol. 2007;7:1–13.
Diaz F, Allan CW, Matzkin LM. Positive selection at sites of chemosensory genes is associated with the recent divergence and local ecological adaptation in cactophilic Drosophila. BMC Evol Biol. 2018;18:144.


Communicating editor: A. Sethuraman