General and Comparative Endocrinology 224 (2015) 84–95

Contents lists available at ScienceDirect

General and Comparative Endocrinology

journal homepage: www.elsevier.com/locate/ygcen

The power of next-generation sequencing as illustrated by the neuropeptidome of the crayfish *Procambarus clarkii*

Jan A. Veenstra *

INCIA UMR 5287 CNRS, Université de Bordeaux, Pessac, France

ARTICLE INFO

Article history:
Received 25 April 2015
Revised 29 May 2015
Accepted 3 June 2015
Available online 3 July 2015

Keywords:
Crustacean female sex hormone
Crustaceans
Venus kinase receptor
Neuropeptide
GPCR
Neuroparsin

ABSTRACT

Transcriptomes of the crayfish *Procambarus clarkii* were analyzed for the presence of transcripts encoding neurohormones, neuropeptides and their receptors. A total of 58 different transcripts were found to encode such ligands and another 82 for their receptors. A very large number of the neuropeptide transcripts appeared to be complete and for those that were not only small parts seemed to be lacking. Transcripts for the neuropeptide GPCRs as well as for the putative receptors for insulin, neuroparsin and eclosion hormone were often also complete or almost so. Of particular interest is the presence of three different neuroparsin genes and two putative neuroparsin receptors. There are also three pigment dispersing hormones as well three likely receptors for these neuropeptides. CNMamide, calcitonin, CCRFamide, natalisin, trissin and relaxin appear to be new crustacean neuropeptides. The recently identified crustacean female sex hormone was also found and in the crayfish appears to be not only expressed in the eyestalk, but in the ovary as well (though not in the testis). Interestingly, there are two other proteins in the crayfish with a structure similar to crustacean female sex hormone, that could be precursors of neurohormones, but these are not expressed by the ovary. The ovary also appears to contain significant numbers of transcripts encoding pigment dispersing hormones, CNMamide as well as glycoprotein B5, but not glycoprotein A2.

1. Introduction

Recombinant DNA technology has revolutionized biology. Whereas manual sequencing allowed the determination of relatively small recombinant DNA sequences, fluorescent labeling and automation made it possible to determine whole genome sequences by the year 2000. The first sequenced genomes were relatively small but a draft for the entire human sequenced followed very quickly. Next-generation sequencing (NGS) has once again moved the limits of feasibility and can be used to sequence entire genomes. However, this remains a challenge when the size of the genome is large and when the species to be sequenced is small and/or when no homozygous individuals are available (see *e.g.* Richards and Murali, 2015).

The largest genome sequenced to date is that of the migratory locust *Locusta migratoria* and this sequence represents a major milestone (Wang et al., 2014). Although current improvements in NGS technology will no doubt make such large genomes easier to sequence, in many cases one is more interested in the coding and non-coding RNAs generated by a genome than in the genomic sequences per se. The NGS technologies have been phenomenal in determining which RNAs are expressed by the genome, whether in specific tissues or in entire individuals. Not surprisingly, it is extensively used for a large number of different projects. Many of these concern the expression of every gene under different experimental conditions in a particular tissue, while the objective of other studies is to determine protein sequences from different species and use those for the construction of phylogenetic trees, such as those that were used to generate such a tree for Arthropods (Misof et al., 2014). As a result there is a enormous amount of data that is often only very partially exploited. For the comparative endocrinologist this allows the exploration of unprecedented amounts of DNA sequences, which has allowed us to document the presence of numerous neuropeptides in species that have never been used before in endocrinological research (*e.g.* Christie et al., 2008, 2010; Christie, 2014a,b). In some cases such data may lead to interesting experiments on those species, but in many cases this will probably not be the case. Although such data is very interesting and valuable, it is of haphazard nature in that it does not supply the entire neuropeptidome of a single species such as a complete genome sequence may provide. It is more interesting to deduce neuropeptidomes from species that have a large genome and are

* Address: INCIA UMR 5287 CNRS, Université de Bordeaux, allée Geoffroy St Hillaire, CS 50023, 33 615 Pessac Cedex, France.
E-mail address: jan.veenstra@u-bordeaux.fr

http://dx.doi.org/10.1016/j.ygcen.2015.06.013
0016-6480/

used in biological research. Studies on several crustaceans have yielded significant numbers of neuropeptide transcripts (*e.g.* Christie, 2014c), but many of these transcripts are incomplete and one suspects that these neuropeptidomes also remain incomplete. We here describe the neuropeptidome of the crayfish *Procambarus clarkii*. The genome of this species has been estimated as 6.2 pg (Bachmann and Rheinsmith, 1973), which would correspond to about 6064 Mbp and this crustacean is commonly used as a research model for neuropeptide research (*e.g.* Yasuda et al., 1994, 2004; Nagasawa et al., 1996; Yasuda-Kamatani and Yasuda, 2000, 2004, 2006). Much to our surprise, even though there are only seven short read archives (SRAs) and two transcriptome shotgun assemblies (TSAs) at NCBI for this species (Tom et al., 2014; Shen et al., 2014; Manfrin et al., 2015), they contain sufficient information to get a very complete picture of the neurohormones, neuropeptides and their receptors encoded by this very large genome.

2. Materials and methods

Two *P. clarkii* TSAs (GBEV00000000.1 and GARH00000000.1) and seven SRAs (SRR870673, SRR1144630, SRR1144631, SRR1265966, SRR1509456, SRR1509457 and SRR1509458) were downloaded from NCBI, while an additional TSA file was graciously made available by Dr. Huaishun Shen. Dr Shen also supplied the crude RNAseq data from the hepatopancreas described by him and his colleagues (Shen et al., 2014). The fasta sequences were extracted from the SRA's with the SRA toolkit (http://www.ncbi.nlm.nih.gov/Traces/sra/?view=software) and all fasta files were made into BLAST databases using the BLAST program (Altschul et al., 1997; Camacho et al., 2009) that were then searched using the BLAST program with the protein sequences of neuropeptides, neurohormones and their receptors. Artemis (Rutherford et al., 2000) was helpful in checking the various sequences for indels and correcting those while Trinity (Haas et al., 2013) was used to improve the transcripts as well as to find additional sequences. There are often SNPs in the sequences that interfere with the production of a single transcript. For example, the two assembled transcripts GBEV01096894.1 and GBEV01064471.1 have an overlap of 78 nucleotides, however a single C–T difference between these sequences leads to two assembled transcripts instead of one. Such sequences have been combined into a single transcript. All final predicted protein sequences as well as the nucleotide sequences encoding them are presented in a Supplementary excel file.

Signal peptides were predicted with Signal P 4.0 (Petersen et al., 2011) and a single signal anchor with Signal P 3.0 (Bendtsen et al., 2004). Convertase cleavage sites prediction was guided by the rules described for insects (Veenstra, 2000).

In order to find putative ligands for identified GPCRs, a simple phylogenetic tree was made with the *Procambarus* GPCRs and those from the termite *Zootermopsis nevadensis*, the most basal insect species for which the GPCRs have been curated and annotated (Veenstra, 2014), together with the protostomian neuropeptide GPCRs that have been deorphanized, the majority of which are from *Drosophila melanogaster*. We only used the concatenated transmembrane regions of the GPCRs for making these trees. For sequence alignment we used clustal omega (Sievers et al., 2011), while Seaview (Gouy et al., 2010) was used for viewing and correcting the alignments and FastTree (Price et al., 2010) for making the tree. This tree is in Supplementary data. Smaller trees of subsets of GPCRs were made including GPCRs from additional species. These trees were made using PhyML (Guindon and Gascuel, 2003) with 100 bootstraps.

We counted the number of individual RNAseq reads in each of the seven SRAs that corresponded to each of the identified neuropeptide precursor and receptor transcripts. For this we only used the DNA sequences that coded for the precursor and did not include the 5′ or 3′ untranslated regions of their putative mRNAs, since in the absence of a genome one cannot exclude the possibility that the Trinity program has joined transcripts from different genes. One should realize that the numbers obtained should be considered semi-quantitative only, but at least they give some impression of where these genes are expressed.

3. Results and discussion

A total of 58 putative neuropeptide precursors were identified (Fig. 1), including a number of recently identified arthropod peptides such as trissin (Ida et al., 2011), EFLamide (Veenstra et al., 2012), natalisin (Jiang et al., 2013), CNMamide (Jung et al., 2014), calcitonin (Veenstra, 2014) and CCRFamide (Conzelmann et al., 2013) as well as the crustacean female sex hormone (Zmora and Chung, 2014). In the absence of a genome it is impossible to know how complete this list is, but it is notable that very few known arthropod neuropeptides are missing. Peptide sequences not encountered are allatotropin, achatin, DENamide, ecdysis triggering hormone and orcokinin B, as well as some neuropeptides that appear limited to spiders (Veenstra, 2015). That no precursor for ecdysis triggering hormone was found, while two putative ETH receptors were identified, is perhaps not surprising as in insects and ticks this neuropeptide is produced exclusively in peripheral neuroendocrine cells (Roller et al., 2010) and it is likely that those cells were not included in the various RNA samples. Perhaps for similar reasons, we neither found the insulin-like factor from the androgenic gland. The gut was similarly not included and hence neuropeptides that are expressed predominantly or exclusively in the gut were likely also missed. This might concern two peptide precursors that in insects seem to be specific for the midgut, the orcokinin-B transcript (Sterkel et al., 2012; Veenstra and Ida, 2014; Chen et al., 2015) and a second transcript from the calcitonin gene (Veenstra, 2014). As in *Daphnia* the orcokinin gene is predicted to produce a transcript encoding a single orcokinin B, it is possible that such a transcript also exists in the crayfish. Although both allatotropin and DENamide genes are present in *Daphnia* (Dircksen et al., 2011), to our knowledge such peptides have never been found in decapodes. Hence their absence from the *Procambarus* transcriptome may reflect the genuine absence of such peptides in this species. In the case of allatotropin there was neither evidence for the presence of an allatotropin receptor, even though many other neuropeptide GPCRs could be identified in the RNAseq data (as no DENamide receptor has been identified so far a similar analysis for this peptide is not possible).

Although the assembly of the majority of the neuropeptide precursors was straight forward, there are a number of limitations inherent to the use of small DNA sequences to assembly consensus sequences. One of them are single nucleotide polymorphisms (SNPs) that can cause the Trinty program to block further extensions of a sequence. In a few cases SNPs were observed between cDNAs encoding the same protein that were derived from either one of the two TSAs. A more serious problem is encountered when there are short repetitive sequences, such as those of neuropeptides encoding a number of closely related or even identical neuropeptides. Indeed, the precursors of allatostatin A, EFLamide, FMRFamide, leucokinin, NPLP-1, orcokinin and pyrokinin present in the two TSAs all had this problem. Such problems were not found in the allatostatin B, natalisin, tachykinin or sulfakinin precursors, where the paracopies are either lower in number, more dissimilar or further separated between one another on the