BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 3 2014, pages 310–316
doi:10.1093/bioinformatics/btt570

Genome analysis

Advance Access publication October 1, 2013

Spatial regulation dominates gene function in the ganglia chain

Dror Hibsh[1,2,3], Hadas Schori[2,3], Sol Efroni[1,*] and Orit Shefi[2,3,*]

[1]Faculty of Life Sciences, [2]Faculty of Engineering and [3]Institute of Nanotechnologies and Advanced Materials, Bar Ilan University, Ramat Gan, Israel 52900

Associate Editor: Martin Bishop

ABSTRACT

Motivation: To understand the molecular mechanisms of neurons, it is imperative to identify genomic dissimilarities within the heterogeneity of the neural system. This is especially true for neuronal disorders in which spatial considerations are of critical nature. For this purpose, *Hirudo medicinalis* provides here an ideal system in which we are able to follow gene expression along the central nervous system, to affiliate location with gene behavior.

Results: In all, 221.1 million high-quality short reads were sequenced on the Illumina Hiseq2000 platform at the single ganglion level. Thereafter, a *de novo* assembly was performed using two state-of-the-art assemblers, Trinity and Trans-ABySS, to reconstruct a comprehensive *de novo* transcriptome. Classification of Trinity and Trans-ABySS transcripts produced a non-redundant set of 76 845 and 268 355 transcripts (>200 bp), respectively. Remarkably, using Trinity, 82% of the published medicinal leech messenger RNAs was identified. For the innexin family, all of the 21 recently reported genes were identified. Spatial regulation analysis across three ganglia throughout the entire central nervous system revealed distinct patterns of gene expression. These transcriptome data were combined with expression distribution to produce a spatio-transcripto map along the ganglia chain. This study provides a resource for gene discovery and gene regulation in future studies.

Contact: orit.shefi@biu.ac.il or sol.efroni@biu.ac.il

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 30, 2013; revised on September 3, 2013; accepted on September 25, 2013

1 INTRODUCTION

Following the complete set of transcripts is elemental for uncovering novel molecular components, and for studying physiological and pathological conditions. Next-generation sequencing technologies provide cost-effective means for quantifying the transcriptome. Recent technological advances have revolutionized the field, making it possible to analyze the transcriptome of selected sets of cells within an organism, leading to the collection of data at a high spatial-transcriptomic resolution. Exploring gene expression profiles quantitatively in combination with the cellular physical distribution in an intact model may provide the link between genomic information, contextual function and behavior. In the brain, associating gene regulation of collections of neurons within their spatial organization and role is of special importance to the study of health and disease. Particularly it is important in neurodegenerative diseases where loss of function is highly correlated with abnormalities in form (Gaggelli *et al.*, 2006; Kanaan *et al.*, 2012; Seixas *et al.*, 2012).

To perform such mapping, the nervous system of the medicinal leech, *Hirudo medicinalis* was selected. This system, a favorable model for investigating development, regeneration and repair, provides a useful model due to its simple structure and distinct behaviors (Coggeshall and Fawcett, 1964). The central nervous system (CNS) of the leech consists of a chain of ganglia that has been characterized anatomically and physiologically, and presents high similarity between ganglia (Coggeshall and Fawcett, 1964). Each such ganglion contains 400 neurons, most with known function and connections (Macagno, 1980; Meriaux *et al.*, 2011). Previous gene expression studies on leech CNS have resulted in the characterization of specific genes (Harvey *et al.*, 1986; Wysocka-Diller *et al.*, 1989). Attempts have been made to clone genes of interest via the candidate gene approach (Blackshaw *et al.*, 2004; Vergote *et al.*, 2004, 2006). Recently, an expression sequence tag database was constructed and is now available to the scientific community (Macagno *et al.*, 2010). Yet, functional genomic studies in the *H.medicinalis* are in their infancy (Kandarian *et al.*, 2012; Macagno *et al.*, 2010).

Recent approaches in RNA-sequencing (RNA-seq) analysis accommodate low amounts of total RNA, multiple condition comparisons and overcome the *de novo* assembly challenge (Birol *et al.*, 2009; Grabherr *et al.*, 2011; Robertson *et al.*, 2010; Sadamoto *et al.*, 2012; Simpson *et al.*, 2009; Vijay *et al.*, 2013; Zhao *et al.*, 2011). Here, an RNA-seq method is used to study transcription along the leech CNS. The transcriptome of three ganglia is examined: close to the head, mid-body and close to the tail. This choice of three representative ganglia enables querying of gene expression profiles and linking gene expression with CNS spatial organization, at the single ganglion level. The choice of three biologically distinct conditions and not the common two-condition differential analysis (Leng *et al.*, 2013) adds a level of complexity, enabling the identification of latent patterns of expression.

In the current study, a transcriptome of the adult leech *H.medicinalis* CNS was produced. A non-redundant set of transcripts has been identified and analyzed in various forms, including gene–isoform relationships, functional relationships, full-length analysis and gene ontology. To verify the full-length analysis and coverage level, the innexin family, which has been previously studied in the leech, was tested (*H.medicinalis* and *Hirudo Verbana*) (Anava *et al.*, 2009; Dykes and Macagno, 2006; Dykes *et al.*, 2004; Kandarian *et al.*, 2012; Phelan, 2005;

[*]To whom correspondence should be addressed.

Yen and Saier, 2007). A differential expression analysis over the three spatially distinct regions has been used to identify patterns of gene expression. The gene expression profiles were characterized and classified into nine major patterns, demonstrating dominant expression distributions along the ganglia chain.

In the present study, advantage is taken of a simple model, the leech CNS, together with a novel differential expression approach, to combine the transcriptome with the spatial configuration, thus, producing a spatio-transcripto map of the CNS. Resultant data provide a highly useful functional and genomic resource for future systemic studies. Moreover, the strategy for transcriptome analysis presented here may be helpful in other similar transcriptome studies.

2 METHODS

2.1 Leech RNA

As described in the Workflow (Supplementary Fig. S1), RNA has been extracted from three adult medicinal leeches (Supplementary Fig. S2). All were obtained from a *H.medicinalis* colony grown in France at Ricarimpex Farm. Leeches are maintained in our animal facility in tanks populated with about 20 leeches in a controlled environment, at 16°C and 12 h/12 h day/night cycle. Before use, leeches were placed on ice for 30 min and then dissected dorsally. Three ganglia were harvested from each leech (numbers 2, 10 and 19) (Fig. 1 and Supplementary Fig. S2). For technical replicas, ganglion number 10 was harvested from three additional leeches, pooled together for RNA isolation and separated into three samples for RNA-seq.

2.2 RNA isolation and quality control

Total RNA was extracted from each ganglion using RNeasy Lipid Tissue (Qiagen). The quality and quantity of each RNA sample was assessed by Agilent's 100 Bioanalyzer pico chip. RNA samples contained a concentration of 45–170 pg/μl.

2.3 RNA amplification

RNA has been amplified using Ovation Kit v2.0 (NuGEN). Before amplification, all samples were lyophilized using a SpeedVac instrument and then suspended in 5 μl of nuclease-free water.

2.4 Illumina sequencing and quality control

Complementary DNAs (cDNAs) were quantified using Nanodrop and Bioanalyzer DNA 1000 Chip (Supplementary Table S1). Two micrograms (in 100 μl) of cDNAs were fragmented using Bioruptor instrument with three 10 s ('on') cycles of sonication interrupted by 90 s pauses ('off'). The library preparation proceeded with the 'END-REPAIR' reaction from the NEB kit (NEBNext) and then with TruSeq DNA/RNA library preparation. cDNA libraries were loaded on a high-sensitivity chip and quantified on the QuBIT instrument (Supplementary Table S2), to prepare the two 6-plex pools that were separated into two pools (Supplementary Fig. S2). The two pools were quantified (molarity) on Bioanalyzer with the High Sensitivity DNA kit and diluted. The cDNA libraries were generated using messenger RNA-seq (mRNA-seq) assay for transcriptome sequencing on Illumina Hiseq2000. Three cDNA libraries were generated from the total RNA of ganglion number 19 and three cDNA libraries were generated from the pooled total RNA of ganglion 10 in equal amounts, and sequencing was performed in one lane to generate 50 bp single end (SE) reads. A similar procedure was carried out for ganglia numbers 2 and 19. Library construction and

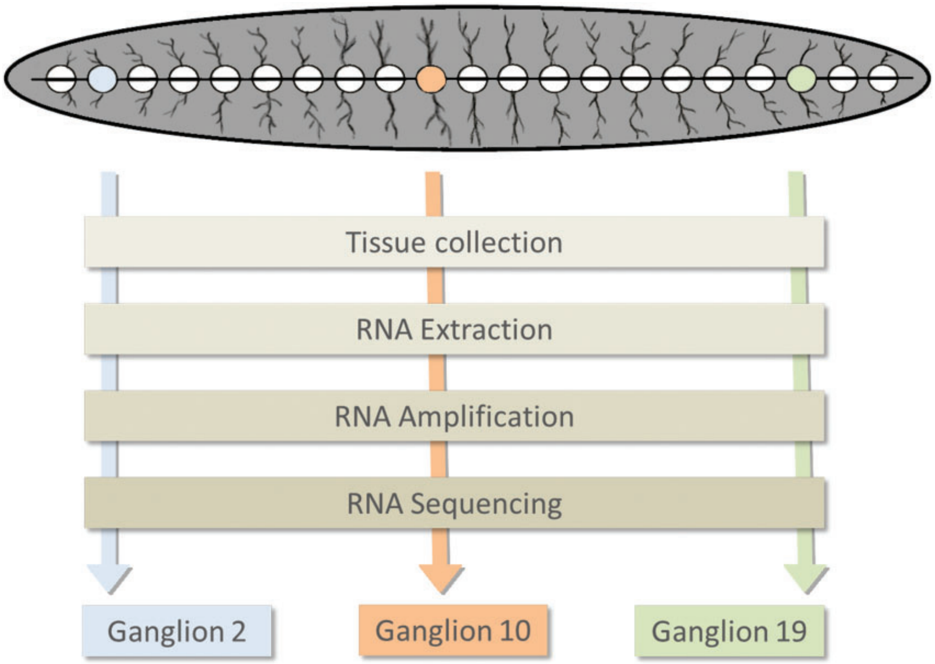


Fig. 1. RNA sample preparation procedure. RNA was collected from three different ganglia. Blue depicts ganglion number 2, orange depicts ganglion number 10 and green depicts ganglion number 19. All ganglia were handled using the same procedures before sequencing. Starting with tissue collection, followed by RNA extraction, then RNA amplification and finally RNA sequencing. Three biological replicates were extracted from each of the three different ganglia (2, 10, 19) and three more samples were taken for technical replicates from ganglion 10 (for more information see Section 2 and Supplementary Fig. S1–S3)

sequencing was performed by a commercial service provider (IGA, Applied Genomics Institute). The sequenced data generated in this study were deposited at Gene Expression Omnibus (record GSE45569 or in Sequence Read Archive (SRA) accession number SRR799260-71).

2.5 *De novo* assembly

Various programs for *de novo* assembly of the 50 bp SE sequence reads were tested to generate a non-redundant set of transcripts/transcripts. Among the various programs available, the publicly available program Trinity (version trinityrnaseq_r2012-03-17) (Grabherr *et al*., 2011) was used. Trinity has been developed for assembly of short reads using de Bruijn graph algorithm by single k-mer. Trinity was executed in the inchworm method and default assembly parameters were used. The publicly available program, Trans-ABySS (version 1.3.2) (Robertson *et al*., 2010), which has been developed for assembly of short reads using de Bruijn graph algorithm by multiple k-mer, was used as well. It has been suggested (Robertson *et al*., 2010) to use the assembly of ABySS (version 1.3.2) (Ning Leng *et al*., 2012) followed by Trans-ABySS. Assembly of transcripts generated by ABySS into transcripts using Trans-ABySS with default parameters for SE sequence reads and k-mer values of (25–50) was performed.

2.6 Mapping reads and estimating counts

The Bowtie algorithm (Version 0.12.7) (Langmead *et al*., 2009) was used to map short reads to the reconstructed transcriptome. After mapping read counts were estimated by two state-of-the-art tools. First, RNA-Seq by Expectation-Maximization (RSEM) (Version 1.1.21) (Li and Dewey, 2011), and then, as recommended by the Cufflinks pipeline (Version 2.0.2) (Roberts *et al*., 2011; Trapnell *et al*., 2010; Trapnell *et al*., 2012), Tophat (Version 2.0.4) (Langmead *et al*., 2009; Trapnell *et al*., 2009) for splice junction mapping, followed by estimating read counts using cufflinks. To obtain a reasonably sized set of transcripts, a smaller set was defined by filtering out transcripts estimated by RSEM with mean counts of <3 and >5000 across all samples. This range based on the *x*–*y* correlation as presented in the scatter plots (Supplementary Fig. S3) has led to ~20% of filtered out transcripts. Those transcripts were irregular, such as ribosomal RNAs or polymerase chain reaction duplicates, products of amplification.

2.7 Full-length and coverage analysis

Identifying whether reconstructed transcripts are full-length is notoriously difficult. For *de novo* assembly, the task is even harder, as there is no reference to work with. Usually for *de novo* assembly, the alternative for organism transcripts is a close organism transcript. This was accomplished here in a few steps. First, published genes were examined and then innexin genes were analyzed as a well-studied family in the leech (Kandarian *et al*., 2012). Blast results of Trinity and Trans-ABySS assemblies were used to detect innexin genes in our reconstructed assemblies. Innexins with an *E*-value cutoff set to e^{-10} were chosen first, after which, for maximizing confidence, only the innexin genes with an *E*-value of 0.0 were chosen, filtered for a perfect match, otherwise the innexins may be aligned to >1 transcript (innexins show high levels of similarity among themselves). For further analysis, only innexin genes that both Trinity and Trans-ABySS reconstructed were used. For the coverage analysis, the same innexin genes reconstructed by both Trinity and Trans-ABySS were used and then Bowtie was used for mapping the raw reads to the innexins.

2.8 Gene–isoform relationship

De novo assembled transcriptomes present a unique challenge in obtaining an accurate gene–isoform relationship. Therefore, the RSEM (Li and Dewey, 2011) tool 'rsem-generate-ngvector' was used to cluster isoforms based on measures directly relating to read mapping ambiguity. This tool first calculates the 'unmappability' of each transcript—the ratio between the number of k-mers with at least one perfect match to other transcripts and the total number of k-mers of this transcript where k is a parameter. Then, an Ng vector is generated by applying a K-means algorithm to the 'unmappability' values with the number of clusters set to 3. This ensures that the mean 'unmappability' scores for clusters are in ascending order. All transcripts with lengths <k are assigned to cluster 3.

2.9 Functional relationships and gene ontology

To establish functional and evolutionary relationships, BLASTX (version 2.2.23) was used to identify sequence conservation. Both Trinity and Trans-ABySS assemblies were run against Swiss-Prot and non-redundant databases downloaded from the National Center for Biotechnology Information (NCBI). Best hits were used as DAVID (Jiao *et al*., 2012) input for detecting enriched gene ontology for the assemblies. Results of the single best hit were extracted and hits with an *E*-value of e^{-10} were considered significant.

2.10 Differential expression

Two approaches were implemented to determine the differential expression of the transcripts. Initially, EBSeq (Version 1.1.3) (Leng *et al*., 2013), which uses an empirical Bayesian approach to accommodate multiple condition comparisons was used. EBSeq uses RSEM counts as input with upper quartile normalization. EBSeq classified transcripts into five patterns with a false discovery rate (FDR) threshold of 0.05 (Fig. 2C). Each pattern was further divided into sub patterns by changing $\neq$ to $\geq$ and $<$ symbols to achieve a more accurate representation of EBSeq patterns (Fig. 2D). Next, EBSeq patterns were divided into three groups by a simple condition: all transcripts in the same state of expression in the first two ganglia were defined in the same group (Fig. 3). In addition, Cuffdiff (Version 2.0.2) (Trapnell *et al*., 2012) was used with upper quartile normalization and with an FDR threshold of 0.05. Cuffdiff uses Cufflinks counts as input for the beta negative binomial approach. Unlike EBSeq, Cuffdiff sorts out only two conditions at a time.

3 RESULTS

3.1 Sequencing and quality control

An amount of 221 138 673 sequence reads were generated, each 50 bp in length, encompassing roughly 11 giga bases (Supplementary Table S3). The expression levels of the technical replicates, estimated by RSEM, showed a strong Pearson correlation (r = ~0.9) (Supplementary Fig. S3). All samples passed FastQC basic statistics for quality estimation, per base sequence quality, per sequence quality score and length distribution.

3.2 *De novo* transcriptome assembly

The *de novo* assembly of the *H.medicinalis* CNS transcriptome was reconstructed by various assembly programs. As described in Section 2, assembly was performed first using Trinity, which generated 76 845 non-redundant transcripts/contigs (>200 bp in length). From the transcripts obtained by Trinity assembly, a set of 22 969 transcripts presented significant alignment scores to the Swiss-Prot database and a set of 26 800 transcripts produced significant alignment to the NCBI non-redundant database. The procedure for the Trans-ABySS assembly was also followed, resulting in a total number of 268 355 non-redundant transcripts/transcripts (>200 bp in length). From the 268 355 transcripts, a set of 85 833 transcripts produced significant alignment to the

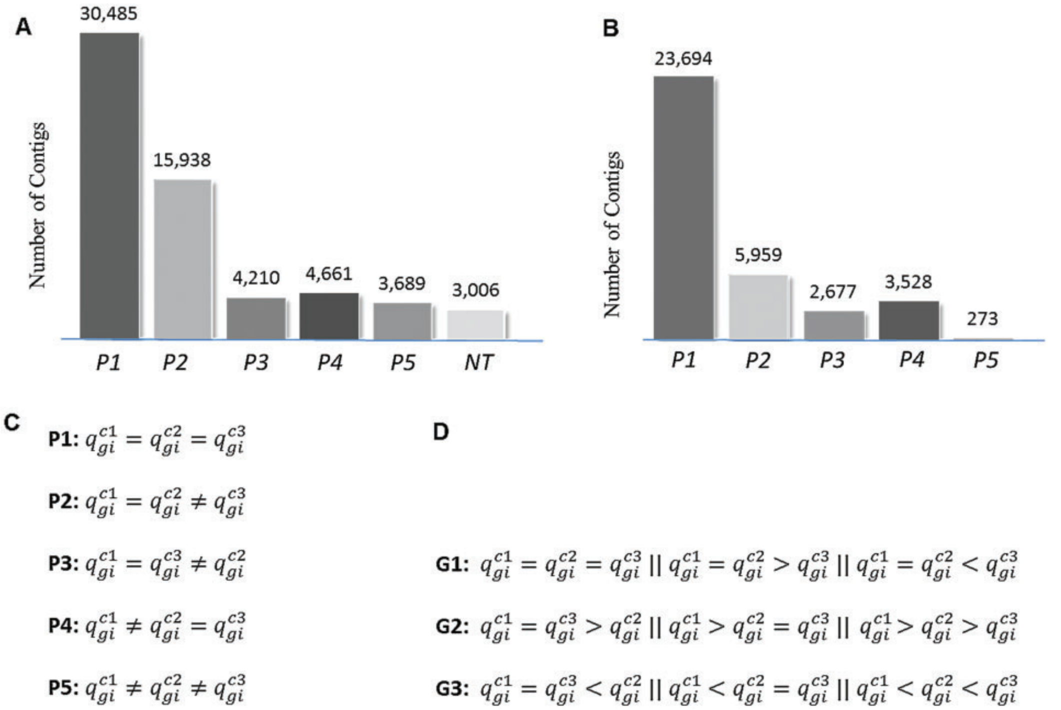


Fig. 2. Classification of Trinity transcripts into patterns and groups. (**A**) Using EBSeq, each of the Trinity contigs were classified into a pattern. In all, 76 845 transcripts were measured across all patterns combined (including those with NoTest (NT) symbol that tags missing expression values). The *x*-axis describes pattern marks, such as P1 for pattern1, P2 for pattern2 and so on. (**B**) Filtering of EBSeq results by setting an FDR threshold of 0.05 yields an updated number of contigs for each pattern. (**C**) Original conditions for classification of contigs into patterns by EBSeq. EBSeq used three conditions (C1, C2, C3) to classify the whole transcript into five possible expression patterns (P1, . . . ,P5). The expression of gene 'i' in c1 is q_{gi}^{c1} and so on. C1, ganglion 2; C2, ganglion 10; C3, ganglion 19. (**D**) Formal representation of conditions defined for the three groups (for more details see Section 2)

Swiss-Prot database and a set of 100 601 transcripts produced significant alignment to the NCBI non-redundant database. Using BLASTN, Trinity was found to have identified 86 of 104 (82%) reported mRNAs, whereas Trans-ABySS identified 82 (78%). The N50, median, average length, total length of transcriptome and GC content for transcripts generated by Trinity was 2188, 550, 1124, 86 436 165 bp, 41.60%, whereas those generated by Trans-ABySS was 1140 , 535 , 809, 217 129658 bp, 42.24% (Supplementary Table S4). The same tests have been performed for *Helobdella robusta* and *Capitella teleta*, as described in Supplementary Table S4. Estimating mapping rate of the assemblies using Bowtie revealed that ~50% of the reads with at least one reported alignment for Trinity assembly and ~30% for Trans-ABySS assembly (Supplementary Table S5).

3.3 Gene–isoform relationship

Using RSEM over Trinity output isoforms showed that most transcripts had 1 isoform, 3 isoforms and 2 isoforms (50 548, 21 273, 5024, respectively), and Trans-ABySS showed that most transcripts had 3 isoforms, 1 isoform and 2 isoforms (212 543, 29 226, 26 586, respectively) (Supplementary Fig. S4).

3.4 Contribution of specific samples to the analysis

By estimating the percentile of transcripts expressed in each sample (Supplementary Fig. S5), we found that according to the Trinity assembly, the highest number of the transcripts expressed across all samples was in sample number 5, which is ganglion number 10. According to the Tran-ABySS assembly, this was sample number 12, which is also ganglion number 10. The lowest number of transcripts came from sample 1, which is ganglion number 2 for both the Trinity and Trans-ABySS assemblies. Other samples had similar quantities between assemblies.

3.5 Full-length and coverage analysis

For both Trinity and Trans-ABySS assemblies, 21 different innexin genes were found, as reported previously (Kandarian *et al*., 2012). With an *E*-value cutoff of 0.0 (see Section 2), 13 innexin genes were found for Trinity assembly as well as for Trans-ABySS assembly, four of which were found to be highly similar, whereas the other nine presented various lengths in different assemblies. Trinity assembled longer transcripts compared with Trans-ABySS in eight of the nine cases (Supplementary Fig. S6). In addition to full-length analysis, coverage of those 13 innexin genes was also tested and found without preferences to either 5′ or 3′, as could be expected (Supplementary Fig. S7).

3.6 Pattern distribution

After filtering out, transcripts of <3 and >5000 mean counts across all samples, 63 077 transcripts were examined for spatial distribution using EBSeq. EBSeq classified Trinity transcripts

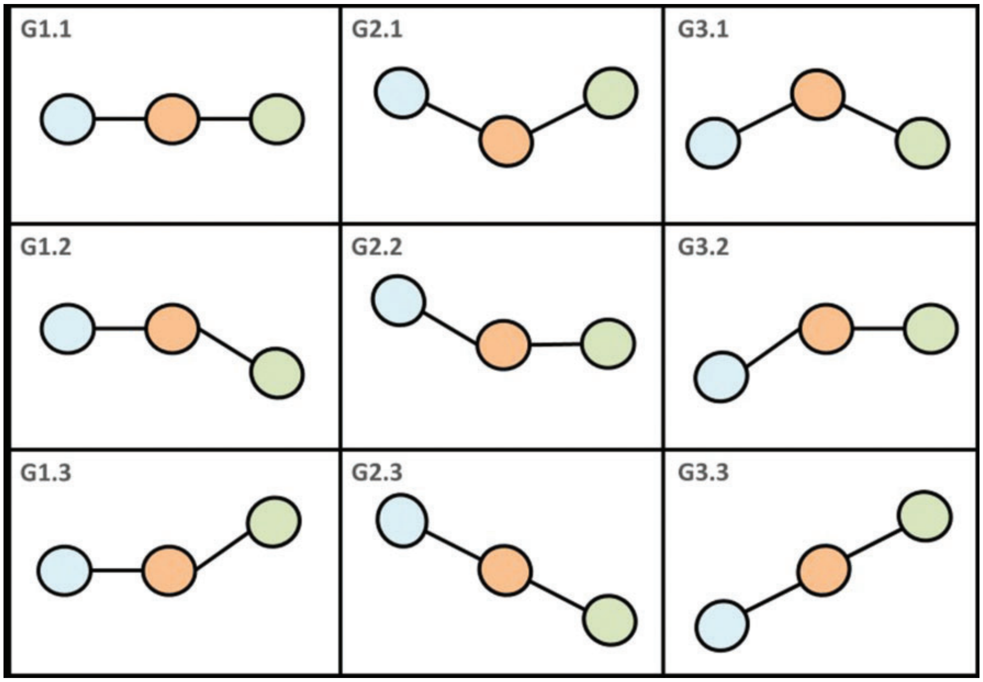
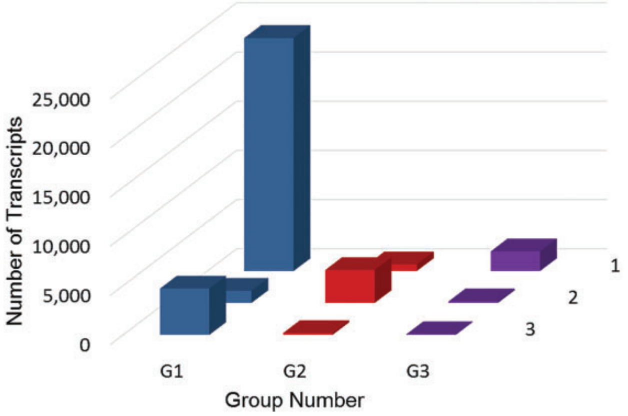


Fig. 3. Definition of sub groups from EBSeq patterns. Each of the matrix elements represents one sub group of the nine that have been classified from EBSeq patterns. Each of these nine elements contains three circles that represent three ganglia. The left circle is ganglion number 2 (blue), the middle circle is ganglion number 10 (orange) and the right is ganglion number 19 (green). EBSeq patterns can be categorized into sub patterns and then divided into three groups (G1, G2, G3) and nine sub groups ('.1', '.2', '.3') as described in Section 2. G1.1 is an example of equal expression between the three ganglia. G2.3 is an example of downregulation between the three ganglia from left to right. G3.3 is an example of upregulation between the three ganglia from left to right

into five patterns, each of which contained a different number of transcripts as shown in Figure 2A. Transcripts were then filtered out by setting an FDR of 0.05, which resulted in less transcripts per pattern (Fig. 2B) and correlated transcripts in each pattern to the BLASTX results. A list of genes expressed within each pattern was found. DAVID analyses showed that many genes identified by EBSeq were involved in neuron projection, nervous system regulation, neurogenesis regulation, neural differentiation, developmental processes and metabolism. Once the groups were defined (as described in Section 2) and each of the sub groups counted, dominant sub groups were identified as described in Figure 4 (G1.1, G2.2, G3.1). By examining this hypothesis using a χ^2 goodness-of-fit test, the data were found in consistent with random distribution (Supplementary Fig. S8A–C). Furthermore, all possible pairings between the leeches were examined (leech number 1 versus 2, 2 versus 3 and 1 versus 3), and the same dominant behavior of sub groups was observed (Supplementary Fig. S9A–C). In addition, using a χ^2 test for independence, a dependency was observed between the states (equal/unequal expression) of the transcripts in ganglion 2 and 10 and those of the transcripts in ganglion 10 and 19 ($\chi^2 = 1851.208$, degrees of freedom (df) = 1, $P < 2.2\text{e-}16$) (Supplementary Fig. S8D). To further establish the statistical significance of the abundance of the identified spatial gene patterns between the nine samples, 30 retagged locations out of the original data were produced (Supplementary Table S7). The identified patterns in the shuffled patterning were reevaluated.



1	23,694	674	2,003
2	1,243	3,361	167
3	4,716	198	75

Fig. 4. Number of transcripts in each group. The *x*-axis represents the group identification [G1, G2, G3]. The *y*-axis represents the number of transcripts found in this state. For example, G1.1 indicates 23 694 transcripts that share equal expression across the ganglia and G3.3 indicates 75 transcripts that share up-gradient expression across the ganglia. Summing the table values is 36 131 and not 76 845 because FDR is set at 0.05

Results from this analysis provide further validation to the statistical tests used to establish the patterns. When Cuffdiff results were classified to identify the EBSeq pattern with a contig match, perfect matches were found (Supplementary Table S8).

4 DISCUSSION

Along with the rapid advances in sequencing technologies, a set of bioinformatics tools has recently been developed for short-read sequence data assembly (Birol *et al*., 2009; Grabherr *et al*., 2011) and analysis (Leng *et al*., 2013; Sadamoto *et al*., 2012; Trapnell *et al*., 2012). *De novo* assembly of short reads without a known reference genome is considered non-standard and non-trivial, but still attracts enough attention to provide the community with an adequate set of tools (Duan *et al*., 2012; Robertson *et al*., 2010; Sadamoto *et al*., 2012; Vijay *et al*., 2013; Zhao *et al*., 2011). Whole-transcriptome sequencing provides a functional genomic view of the studied tissue. Recent developments in high-throughput technologies have provided a plethora of molecular data, but a thorough account of RNA expression levels in the *H.medicinalis* is still lacking.

In this study, a *de novo* assembly of transcriptome is produced using short reads from non-model annelids, the *H.medicinalis*, for which public sequence data are still limited. The production of a verified transcriptome is used here as a tool to study the CNS of *H.medicinalis*. *H.medicinalis* is one of the most important annelids and as such a valuable model for studying nervous system structure, function, development, regeneration and repair. More than 221 million sequence reads were generated for *H.medicinalis*.

Validation of the different assembly tools significantly affects the assembly output and is needed for optimal results. Trinity and Trans-ABySS have been used for the reconstruction of the *H.medicinalis* transcriptome. More reported mRNAs of *H.medicinalis* have been identified using Trinity. Moreover, examination of full-length transcripts reveals that Trinity reconstructed fuller transcripts for the innexin family, making Trinity transcripts the choice for downstream analyses. Furthermore, we found that the N50, median and average lengths of the transcripts generated using Trinity were longer (Grabherr *et al*., 2011; Mende *et al*., 2012) and with higher similarity to related organisms as shown in Supplementary Table S4 (a comparison with Trans-ABySS) (Macagno *et al*., 2010).

A non-redundant set of 76 845 transcripts is reported. The gene–isoform relationship of the *H.medicinalis* transcripts shows that the largest fraction of transcripts presents a single isoform. About 30% of the transcripts show significant homology with Swiss-Prot sequences. Comparison with other openly available leech datasets, the transcriptome of *H.robusta*, shows a 39% identity at the amino acid level (Supplementary Table S6). Interestingly, a comparison with *Capitella capitata* (a polychaete) shows similar high levels of amino acid identity (34%) (Supplementary Table S6). The unannotated transcripts (Chopin *et al*., 2000; Frobius and Seaver, 2006; Murray *et al*., 2004; Salzet *et al*., 2000; Tahtouh *et al*., 2009; Tasiemski *et al*., 2004) might indicate novel, and possibly, species-specific functions. Perhaps most importantly, the expression pattern distribution in the leech nervous system presents a significantly larger class of transcripts that are co-expressed along the leech's spatial configuration. Other patterns are much smaller and do not co-express, but rather dominate specific expression patterns along the nervous system. This may indicate that (i) most of the transcripts tend to be located in the ganglia butare not ganglion specific. For example, HSP90 belongs to the previously described 'pattern 1', which supports our classification patterns, and (ii) within each group there is a sub group more dominant, which may suggest that spatial regulation dominates gene function in the ganglia chain.

In conclusion, the specific genes and their spatial templates of expression across the CNS of a whole organism provide the community, for the first time, a functional affiliation between genomic and nerve loci.

ACKNOWLEDGEMENTS

The authors wish to thank Shahar Alon for fruitful discussions and invaluable advice, Helit Cohen for her help with experimentation, Liran Carmel for his help with analyses and advice and Adva Rodan for graphical assistance. The *H.robusta* and *C.teleta* sequence data were produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/in collaboration with the user community.

Funding: This work was supported (in part) by the EU-FP7 People IRG Marie Curie Grants (239482) (to O.S.) and by the Israel Science Foundation for Individual Research Grants (1403/11) (to O.S.)

Conflicts of Interest: none declared.

REFERENCES

Anava,S. *et al*. (2009) Innexin genes and gap junction proteins in the locust frontal ganglion. *Insect Biochem. Mol. Biol*., **39**, 224–233.

Birol,I. *et al*. (2009) *De novo* transcriptome assembly with ABySS. *Bioinformatics*, **25**, 2872–2877.

Blackshaw,S.E. *et al*. (2004) Identifying genes for neuron survival and axon outgrowth in *Hirudo medicinalis*. *J. Anat*., **204**, 13–24.

Chopin,V. *et al*. (2000) Therostasin, a novel clotting factor Xa inhibitor from the rhynchobdellid leech, *Theromyzon tessulatum*. *J. Biol. Chem*., **275**, 32701–32707.

Coggeshall,R.E. and Fawcett,D.W. (1964) The fine structure of the central nervous system of the leech, *Hirudo medicinalis*. *J. Neurophysiol*., **27**, 229–289.

Duan,J. *et al*. (2012) Optimizing de novo common wheat transcriptome assembly using short-read RNA-Seq data. *BMC genomics*, **13**, 392.

Dykes,I.M. *et al*. (2004) Molecular basis of gap junctional communication in the CNS of the leech Hirudo medicinalis. *J. Neurosci*., **24**, 886–894.

Dykes,I.M. and Macagno,E.R. (2006) Molecular characterization and embryonic expression of innexins in the leech Hirudo medicinalis. *Dev. Genes. Evol*., **216**, 185–197.

Frobius,A.C. and Seaver,E.C. (2006) Capitella sp. I homeobrain-like, the first lophotrochozoan member of a novel paired-like homeobox gene family. *Gene. Expr. Patterns*., **6**, 985–991.

Gaggelli,E. *et al*. (2006) Copper homeostasis and neurodegenerative disorders (Alzheimer's, prion, and Parkinson's diseases and amyotrophic lateral sclerosis). *Chem. Rev*., **106**, 1995–2044.

Grabherr,M.G. *et al*. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol*., **29**, 644–652.

Harvey,R.P. *et al*. (1986) Cloning and expression of a cDNA coding for the anti-coagulant hirudin from the bloodsucking leech, *Hirudo medicinalis*. *Proc. Natl Acad. Sci. USA*, **83**, 1084–1088.

Jiao,X. *et al*. (2012) DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics*, **28**, 1805–1806.

Kanaan,N.M. *et al.* (2012) Axonal degeneration in Alzheimer's disease: when signaling abnormalities meet the axonal transport system. *Exp. Neurol.*, **246**, 44–53.

Kandarian,B. *et al.* (2012) The medicinal leech genome encodes 21 innexin genes: different combinations are expressed by identified central neurons. *Dev. Genes Evol.*, **222**, 29–44.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Leng,N. *et al.* (2013) EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, **29**, 1035–1043.

Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, **12**, 323.

Macagno,E.R. (1980) Number and distribution of neurons in leech segmental ganglia. *J. Comp. Neurol.*, **190**, 283–302.

Macagno,E.R. *et al.* (2010) Construction of a medicinal leech transcriptome database and its application to the identification of leech homologs of neural and innate immune genes. *BMC Genomics*, **11**, 407.

Meriaux,C. *et al.* (2011) Multiple changes in peptide and lipid expression associated with regeneration in the nervous system of the medicinal leech. *PLoS One*, **6**, e18359.

Mende,D.R. *et al.* (2012) Assessment of metagenomic assembly using simulated next generation sequencing data. *PloS one*, **7**, e31386.

Murray,J.C. *et al.* (2004) Endothelial monocyte-activating polypeptide-II (EMAP-II): a novel inducer of lymphocyte apoptosis. *J. Leukoc. Biol.*, **75**, 772–776.

Ning Leng,J.D. *et al.* (2012) *De novo* transcriptome assembly with ABySS. *Bioinformatics*, 2872–2877.

Roberts,A. *et al.* (2011) Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*, **27**, 2325–2329.

Robertson,G. *et al.* (2010) *De novo* assembly and analysis of RNA-seq data. *Nat. Methods*, **7**, 909–912.

Sadamoto,H. *et al.* (2012) *De novo* sequencing and transcriptome analysis of the central nervous system of mollusc Lymnaea stagnalis by deep RNA sequencing. *PLoS One*, **7**, e42546.

Salzet,M. *et al.* (2000) Theromin, a novel leech thrombin inhibitor. *J. Biol. Chem.*, **275**, 30774–30780.

Seixas,A.I. *et al.* (2012) Loss of junctophilin-3 contributes to Huntington disease-like 2 pathogenesis. *Ann. Neurol.*, **71**, 245–257.

Simpson,J.T. *et al.* (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res.*, **19**, 1117–1123.

Tahtouh,M. *et al.* (2009) Evidence for a novel chemotactic C1q domain-containing factor in the leech nerve cord. *Mol. Immunol.*, **46**, 523–531.

Tasiemski,A. *et al.* (2004) Molecular characterization of two novel antibacterial peptides inducible upon bacterial challenge in an annelid, the leech *Theromyzon tessulatum*. *J. Biol. Chem.*, **279**, 30973–30982.

Trapnell,C. *et al.* (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, **7**, 562–578.

Trapnell,C., Pachter,L. and Salzberg,S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.

Trapnell,C. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.

Vergote,D. *et al.* (2004) Up-regulation of neurohemerythrin expression in the central nervous system of the medicinal leech, Hirudo medicinalis, following septic injury. *J. Biol. Chem.*, **279**, 43828–43837.

Vergote,D. *et al.* (2006) Proteome modifications of the medicinal leech nervous system under bacterial challenge. *Proteomics*, **6**, 4817–4825.

Vijay,N. *et al.* (2013) Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. *Mol. Ecol.*, **22**, 620–634.

Wysocka-Diller,J.W. *et al.* (1989) Characterization of a homologue of bithorax-complex genes in the leech Hirudo medicinalis. *Nature*, **341**, 760–763.

Yen,M.R. and Saier,M.H. Jr (2007) Gap junctional proteins of animals: the innexin/pannexin superfamily. *Prog. Biophys. Mol. Biol.*, **94**, 5–14.

Zhao,Q.Y. *et al.* (2011) Optimizing *de novo* transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics*, **12** (**Suppl. 14**), S2.