Gene set enrichment analysis of neuroendocrine system of the silkworm *Bombyx mori*

Gabor Beke[1], Matej Stano[1], Ivana Daubnerova[2], Dusan Zitnan[2], Lubos Klucar[1]

[1]Institute of Molecular Biology, Slovak Academy of Sciences, Bratislava, Slovakia
[2]Institute of Zoology, Slovak Academy of Sciences, Bratislava, Slovakia

Introduction

Silkworm *Bombyx mori* represents a model organism for studying the neuroendocrine system in invertebrates. This system of nervous organs and endocrine glands regulates large number of life functions including movement, ecdysis, courting and mating. Our aim was to perform gene set enrichment analysis on several endocrine glands samples, in order to better understand differences in gene expression of these samples. This is a side project based on RNA-seq data analysis of neuroreceptors and neuropeptides of *B. mori*.

Materials and Methods

Mappings to the transcriptome were performed using the *Bowtie2* aligner with allowance of unlimited multi-mappings for each read. *eXpress* was used to quantificate transcript expression level. The result of *eXpress* were cluestered and analyzed for gene set enrichment. For the gene set enrichment analysis R (v3.1.2) and its additional packages were used:

- R for clusteing the results of *eXpress*,
- *gplots* to visualise clustering results in heatmaps,
- *topGO* for the gene set enrichment analysis and to visualise the most significant gene ontology terms.

Results

We have analysed strand-specific Illumina RNA-seq data from *B. mori* samples originated from different endocrine glands of both sexes and several developmental stages. Reads were mapped to the *B. mori* transcriptome using the *Bowtie 2* aligner. Since the silkworm genome is abundant in repetitive sequences (Mita *et al.*, 2004), the mapping allowing unlimited multi-mappings (required by *eXpress*) was extremly computing demanding. Since *B. mori* genome is not completly sequenced, we have used transcriptom consisting of 43 160 fragments, ranging in size from 43 bp to 55 293 bp (N50 = 1 359 bp). Transcript level RNA-seq quantification was performed using the *eXpress* tool based on an online expectation–maximization algorithm. Mapping statistics is showed in Table 1. Range of FPKM values for each sample is described in Table 2. Obtained data were consequently analysed in several gene set enrichment packages including *topGO* and *gplots* (both R packages), where genes were clustered based on their expression level. Hierarchical clustering was calculated using Pearson and Spearman's correlation. We have visualized results of clustering using heatmaps (Fig. 1.). Based on te heatmaps we have selected clusters with the highest expression level. These highly expressed genes were funcionally analysed, the most significant GO terms were identified in these groups. The 10 most significant GO terms for each GO category presented in all groups are showed in Table 3.

Table 1. Mapping statistics

Single-end Strand-specific	reads	aligned	aligned >1 times
CNS Pharate Pupa	73M	79,32%	14,42%
Brain SG male	73M	76,17%	14,34%
Brain SG female	66M	76,46%	13,88%
PTG male	67M	77,83%	13,67%
PTG female	60M	77,77%	13,67%
AG TAG male	73M	62,86%	10,73%
AG TAG female	84M	69,47%	13,69%
Acc glands male	67M	78,50%	13,68%
Acc glands female	70M	79,59%	13,56%
Testes	71M	80,16%	12,25%
Ovaries	69M	82,34%	25,60%
H organ Pharate	84M	80,35%	14,22%
Paired-end Unstranded			
ca male	25M	21,27%	2,71%
ca female	26M	27,77%	3,53%
Single-end Unstranded			
inka 1	7M	35,72%	9,62%
inka 2	14M	37,82%	10,38%
inka 3	12M	42,30%	11,21%
inka 4	15M	36,28%	11,41%

Table 2. Calculated FPKM range

	FPKM values calculated by *eXpress*		
Single-end Strand-specific	**median**	**avg**	**max**
CNS Pharate Pupa	0,13	23,8	20 939
Brain SG male	0,14	21,7	22 394
Brain SG female	0,13	21,3	19 455
PTG male	0,13	21,6	26 251
PTG female	0,12	21,2	16 732
AG TAG male	0,13	21,1	16 790
AG TAG female	0,13	21,3	32 501
Acc glands male	0,08	29,0	38 544
Acc glands female	0,09	24,1	37 155
Testes	0,14	22,8	27 234
Ovaries	0,11	30,4	62 886
H organ Pharate	0,11	25,2	41 274
Paired-end Unstranded			
ca male	0,00	20,7	34 171
ca female	0,01	20,7	28 253
Single-end Unstranded			
inka 1	0,09	45,4	815 438
inka 2	0,03	23,4	69 519
inka 3	0,01	23,4	75 924
inka 4	0,01	26,0	228 576

Table 3. Top 10 GO terms in clusters

Molecular function	
GO:0000166	nucleotide binding
GO:0003676	nucleic acid binding
GO:0003677	DNA binding
GO:0003700	sequence-specific DNA binding transcription factor activity
GO:0003824	catalytic activity
GO:0004674	protein serine/threonine kinase activity
GO:0005506	iron ion binding
GO:0005509	calcium ion binding
GO:0005515	protein binding
GO:0005525	GTP binding
Biological proces	
GO:0005975	carbohydrate metabolic process
GO:0006355	regulation of transcription, DNA-templated
GO:0006464	cellular protein modification process
GO:0006468	protein phosphorylation
GO:0006508	proteolysis
GO:0006886	intracellular protein transport
GO:0007018	microtubule-based movement
GO:0007165	signal transduction
GO:0007166	cell surface receptor signaling pathway
GO:0007186	G-protein coupled receptor signaling pathway
Cellular Component	
GO:0005615	extracellular space
GO:0005622	intracellular
GO:0005634	nucleus
GO:0005743	mitochondrial inner membrane
GO:0005783	endoplasmic reticulum
GO:0008305	integrin complex
GO:0016020	membrane
GO:0016021	integral component of membrane
GO:0016023	cytoplasmic membrane-bounded vesicle
GO:0030286	dynein complex

Figure 1. Heatmaps of clustered genes based on their expression levels. A: Columns (samples) were clustered by Spearman's correlation, rows (genes) were clustered by Pearson correlation. **B:** Samples were clustered by Perason correlation, genes by Spearman's correlation.

Conclusions

Reads from 18 samples of *B. mori* from different endocrine glands of both sexes and several developmental stages were mapped to the transcroptime. Gene expression level for each gene was estimated. We have clustered samples and genes by their level of expression. Clusters were visualized using heatmaps. On selected clusters gene set enrichment analysis were performed. The most overrepresented GO terms presented in groups were too common, mostly corresponding to the binding activity against wide range of compounds.

Acknowledgements

The study is supported by the APVV-0827-11, VEGA 2/0164/15, IMTS 26230120002 and IMTS 26210120002 grants.

References

Mita K *et al*. (2004) The genome sequence of silkworm, *Bombyx mori. DNA Res.* **11**(1), 27-35. http://dx.doi.org/10.1093/dnares/11.1.27