BMC Research Notes

DATA NOTE Open Access

SpiroESTdb: a transcriptome database and online tool for sparganum expressed sequences tags

Dae-Won Kim[1†], Dong-Wook Kim[2†], Won Gi Yoo[1†], Seong-Hyeuk Nam[2], Myoung-Ro Lee[1], Hye-Won Yang[4], Junhyung Park[5], Kyooyeol Lee[5], Sanghyun Lee[1], Shin-Hyeong Cho[1], Won-Ja Lee[1], Hong-Seog Park[2,3*] and Jung-Won Ju[1*]

Abstract

Background: Sparganum (plerocercoid of *Spirometra erinacei*) is a parasite that possesses the remarkable ability to survive by successfully modifying its physiology and morphology to suit various hosts and can be found in various tissues, even the nervous system. However, surprisingly little is known about the molecular function of genes that are expressed during the course of the parasite life cycle. To begin to decipher the molecular processes underlying gene function, we constructed a database of expressed sequence tags (ESTs) generated from sparganum.

Findings: SpiroESTdb is a web-based information resource that is built upon the annotation and curation of 5,655 ESTs data. SpiroESTdb provides an integrated platform for expressed sequence data, expression dynamics, functional genes, genetic markers including single nucleotide polymorphisms and tandem repeats, gene ontology and KEGG pathway information. Moreover, SpiroESTdb supports easy access to gene pages, such as (i) curation and query forms, (ii) *in silico* expression profiling and (iii) BLAST search tools. Comprehensive descriptions of the sparganum content of all sequenced data are available, including summary reports. The contents of SpiroESTdb can be viewed and downloaded from the web (http://pathod.cdc.go.kr/spiroestdb).

Conclusions: This integrative web-based database of sequence data, functional annotations and expression profiling data will serve as a useful tool to help understand and expand the characterization of parasitic infections. It can also be used to identify potential industrial drug targets and vaccine candidate genes.

Keywords: Sparganum, Plerocercoid, *Spirometra erinacei*, Expressed sequence tags (ESTs), Database

Findings

Sparganum is the plerocercoid larva of the genus *Spirometra erinacei* of pseudophyllidean tapeworms. Human sparganosis occurs occasionally through the ingestion of freshwater contaminated with proceroid-infected cyclops or through contact with plerocercoid-infected animal hosts, such as frogs and snakes. Sparganosis is reported in many countries but is most common in eastern Asia [1]. Ingested spargana have the ability to invade various organs, such as the eye, subcutaneous tissues, abdominal wall, brain, spinal cord, lung, breast, and others [2,3]. Human sparganosis can cause diverse symptoms, such as non-specific irritation, uncertain pain, tissue mass formation, or headache, or it can cause no symptoms that affect the internal organs [4]. Thus, sparganum has become an important parasite for human health, and research into this organism will greatly contribute to the identification and characterization of genes and pathways involved in parasite development and function.

For contemporary functional genomic studies, the significant efforts invested in the creation of integrative transcript databases of expressed sequence tags (ESTs) derived from full-length cDNA libraries have provided opportunities to direct gene discovery and functional analysis. However, in the absence of a fully sequenced genome, parasite transcriptome data have only been analyzed in a limited number of studies [5-8]. Therefore, the motivation for this project was to develop a database

* Correspondence: hspark@kribb.re.kr; junomics@gmail.com
† Contributed equally
[1]Division of Malaria and Parasitic Diseases, Korea National Institute of Health, Osong 363-951, Republic of Korea
[2]Genome Resource Center, Korea Research Institute of Bioscience and Biotechnology (KRIBB), Daejeon 305-806, Republic of Korea
Full list of author information is available at the end of the article

and tool that would provide parasite researchers with flexible access to parasite-specific information about sparganum that is relevant for studying the production of parasite proteins and identifying molecules involved in key biological pathways that might serve as targets for diagnostic markers and treatments for the control of sparganum.

Database contents and construction

To obtain a comprehensive view of sparganum gene activity, we constructed a non-normalized cDNA library derived from parasites isolated from wild snakes. In total, 5,760 randomly selected clones were picked and sequenced. After pre-processing, which included base calling and vector contamination and repetitive element trimming with Seqclean (http://www.tigr.org/tdb/tgi/software), we obtained 5,655 high-quality EST sequences greater than 100 bp in length using Phred score of 13 [9,10]. The sequences reported in this paper have been deposited in the NCBI database, under accession numbers HS514072-HS519705. We then assembled the 5,655 ESTs into 1,787 unigenes (910 contigs and 877 singletons) using the commercial software TGICL [11] and CAP3 [12] with the default parameters. To identify putative homology with known nucleotide and protein sequences, we performed separate BLASTN searches (Query Coverage and protein sequences, we performed separate BLASTN database, under -length cDNAaa, Identity range and protein sequences[13] for the 1,787 unigenes in the NCBI NR non-redundant nucleotide [14], KEGG [15] and UniProt [16] databases. To achieve better classification of the biological function of the unigenes, we also used InterProScan [17], Gene Ontology (GO) analysis [18], Tandem Repeats Finder (TRF) [19] and AutoSNP [20] (Figure 1). There were 1,262 unigenes (71%) shared homology with any other predicted or known molecules in public databases (Table 1).

SpiroESTdb architecture and web interface

SpiroESTdb is a relational database that uses an Oracle 11 g database management system and consists of a total of 22 tables. The database runs on a RedHat Enterprise Linux 5.5 platform with an Apache web server, and it was implemented with JSP (Java Server Pages), Java Servlet technology and the AJAX framework. The web interfaces were designed using HTML with some scripts written in JavaScript, and they use cascading style sheet (CSS) properties. The database is currently designed to work best with Microsoft Internet Explorer 8 (optimal resolution 1024 × 800). SpiroESTdb provides a user-friendly interface with seven main menus enabling access to (1) pre-processing reports, (2) clustering and assembling reports, (3) functional annotation reports, (4) curation for more accurate annotations, (5) expression profiling, (6) BLAST, and (7) a downloadable summary of all raw data and analyzed results.

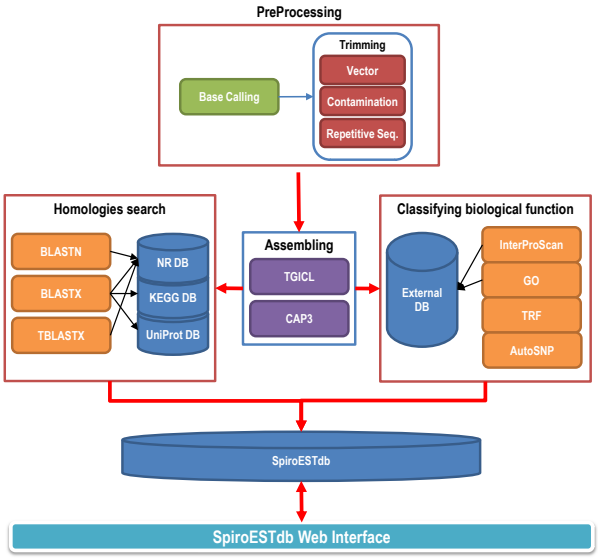

Figure 1 Workflow schema of the pipeline. The architecture of the pipeline includes four components: preprocessing for removing sequence contamination, transcript assembling for unigene clusters, homologies search and classifying biological functions. All data were deposited in SpiroESTdb. The User Interface (UI) supports end-user authentication, the profiling, selection and enactment of workflows.

Utility and discussion

General features of SpiroESTdb components and tools

Preprocessing information

To create the SpiroESTdb database, we employed a semi-automated annotation pipeline that uncovered genomic features in the raw EST sequences assigned for functional annotations based on the BLAST algorithms. First, the pre-processing information for generating high-quality consensus gene sequences is provided as a summary table in the pre-processing and assembling report (Figure 2a). The graphical display of these contigs with alignment sequence information and the trace viewer module allow for the evaluation of the raw sequencing data (Figure 2b and 2c).

Table 1 Summary of sparganum transcriptome

	Numbers
Total sequence reads	5,760
Total analyzed reads	5,655
Total assembled sequences (average size)	1,787 (725 bp)
Contigs	910 (787 bp)
Singlets	877 (661 bp)
Total annotated genes (non-redundant set)*	1,262

*Homologies search was performed by BLASTN, BLASTX, TBLASTX, Uniprot, KEGG and InterProScan

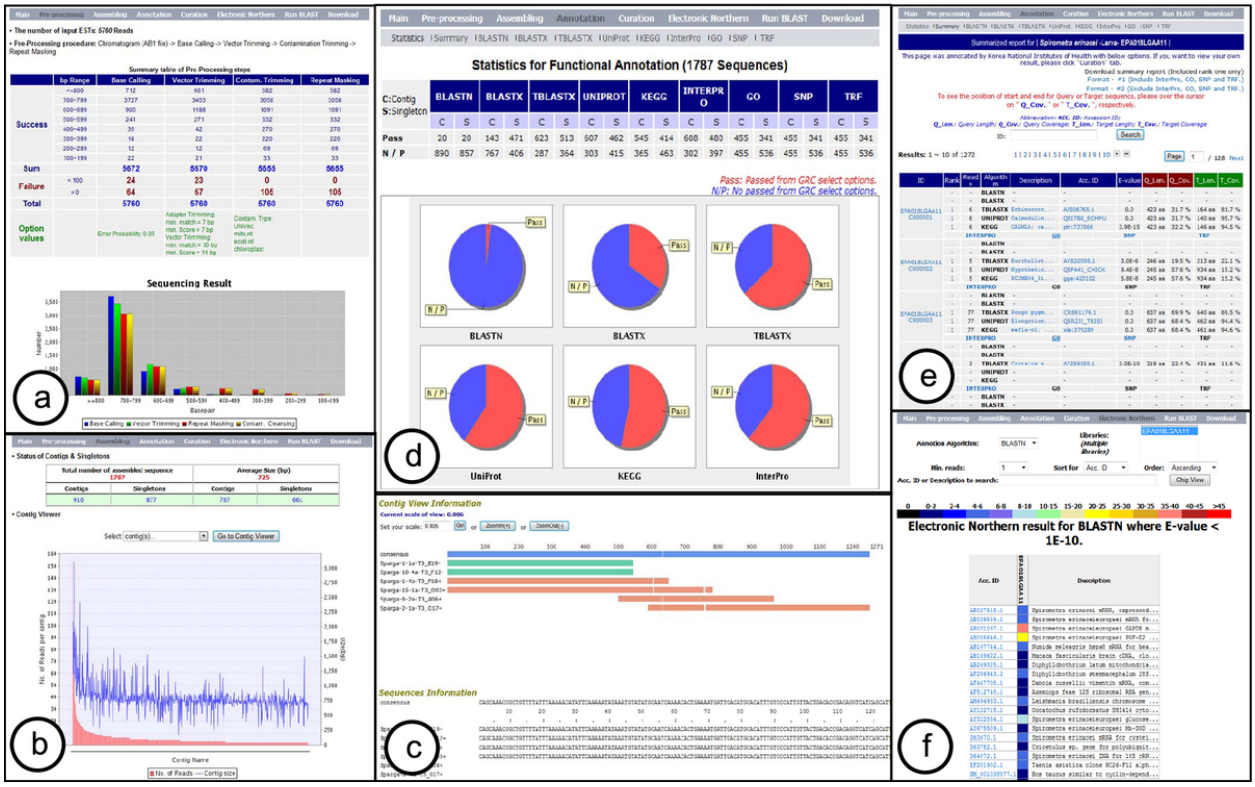

Figure 2 Screenshots of the SpiroESTdb web application. The panels show examples of results generated from general functions. (a) A pre-processing report showing cleansing sequence; (b) An assembly report for viewing contigs and reads; (c) Information of read distribution within a contig; (d) A statistics page describing an annotation result; (e) A summary page of various annotation results; (f) An example of secondary mining information showing *in silico* expression profiling information for all cDNA libraries.

Functional annotation

The online database of SpiroESTdb also provides comprehensive information on the ESTs through the "Annotation" section, which consists of 11 categories including Statistics, Annotation Report, BLASTN, BLASTX, TBLASTX, Uniprot, KEGG, InterPro, GO, SNP and TRF. In particular, in the case of the Statistics page, users can see all of the detailed functional annotation information describing how the unigenes were annotated and taxonomically distributed in the system (Figure 2d). Each category using the BLAST algorithm [13] allows users access to annotation information including unigene IDs and sequence information, annotated unique identifiers, sequence lengths, description, aligned start and end positions, matched similarity, and significant scores that were assigned by the prediction algorithms. SpiroESTdb allows users to view detailed alignment information by clicking on the description of each unigene (Figure 2e). InterProScan [17] is a predictive model for identifying protein domains, families and functional sites using diverse source databases. The InterPro category provides comprehensive information about the functional domains present in previously unidentified genes via a homology search. The potential assignment ID results are integrated into the GO category. As additional data, we also list single nucleotide polymorphisms in the SNP category and copy number variants in the TRF category.

Specific search function

Existing information can be retrieved for any unigene using full-text matching (against the read id, consensus id, gene name, gene accession number and functional description) in the "Annotation" page. With a gene or protein, users are able to blast a query sequence against raw read data, contigs and singletons, via the "Run BLAST" page, to find and visualize potential homologous matches.

Personalized curation service and "Electronic Northern"

For some genes that have not previously been studied, the best hits identified using a homology search will provide ambiguous annotation descriptions such as "predicted protein" or "hypothetical protein". To increase the quality of the BLAST results and to offer biologically meaningful information, SpiroESTdb supports a user-friendly curation service that allows experts to manually

edit functional annotations based on an experimental or heuristic approach. In addition, we programmed the "Electronic Northern" module to detect differential expression between different libraries and/or among unigenes within the same library by comparing accession IDs from BLAST results (Figure 2f).

Practical examples using SpiroESTdb

SpiroESTdb includes a "Electronic Northern" module showing a profile of highly expressed genes that have been evaluated to determine the number of reads contained within one contig. The evaluation values are visualized as a spectrum of colors; black corresponds to the lowest value and red to the highest. Highly expressed genes can be considered novel genes with important biological functions for the parasite's survival and serve as drug targets for sparganosis treatment. For example, users can choose "Electronic Northern" on the sub-main menu and access several options; in the example shown, "TBLASTX" selection in the "Annotation Algorithm" field, the library selection in the "Libraries (Multiple libraries)" field and 30 or over in the "Min. reads" field (Figure 3a). The highly expressed gene with the most reads is a gene encoding *Spirometra erinacei* cytoplasmic antigen containing repeat epitope (U50190.1, 158 reads), followed by fibronectin (FN1, XM_421868.2, 153 reads), *Schistosoma japonicum* clone ZZZ405 mRNA sequence (AY223431.1, 90 reads), *Chlamys farreri* ribosomal protein S19 mRNA (82 reads) and *S. erinacei* mRNA for antigenic polypeptide (AB019222.1, 76 reads) (Figure 3b). In addition to these genes, a number of genes in the energy production pathway in the parasite, such as fructose-bisphosphate aldolase and glyceraldehyde-3-phosphate dehydrogenase, were highly expressed. Of the two antigens found to be highly expressed, one (U50190.1) must still be experimentally validated. FN is a ubiquitous and abundant glycoprotein that represents a combination of three independent domains, FN1, FN2 and FN3. Through interactions with different receptors, FN plays an important role in mediating cellular adhesion and migration processes, including embryonic development and wound healing [21]. Although the function of FN in the parasites is not clearly defined, FN is thought to have various functions that promote parasite survival in the host, such as providing a structural basis for cell adhesion, transducing signals for cell proliferation and apoptosis, and contributing to host defenses [22,23]. In other words, some information assigned by each annotation algorithm can be typed in "Acc. ID or Description to search" such as keyword, partial description and accession number.

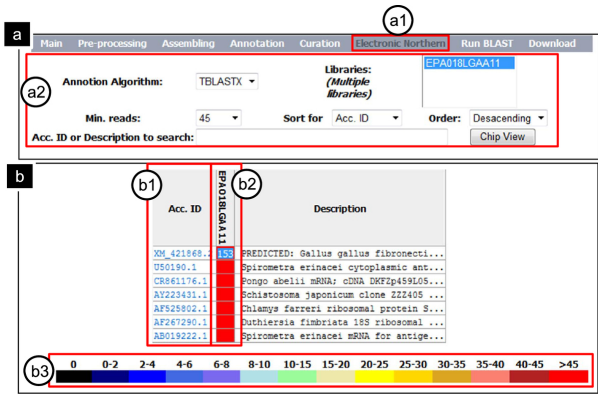

Figure 3 A practical example using SpiroESTdb with "Electronic Northern". (a) The search interface. The steps involved in a "Electronic Northern" search are as follows: (a1) choose "Electronic Northern"; (a2) select search options for generating the appropriate output for your query, including "Annotation Algorithm", "Min.reads" and "Libraries". In the "Libraries" menu, the user can select a specific library; enter any additional information in the search field, if desired; and submit the query. (b) "Electronic Northern" results from a search for highly expressed genes. (b1) By selecting items under 'Acc.ID', detailed information can be obtained from the NCBI protein database. (b2) The number of reads can be displayed by double-clicking the center of the box in the corresponding color. (b3) The color spectrum indicates the expression level (the number of reads, n), ranging from low (black, n = 0) to high (red, n > 45).

Conclusion

Further development of the database will involve updating the experimental records for verified sparganum (plerocercoid of *Spirometra erinacei*) genes with additional functional annotations, such as information about drug targets and vaccine candidates, maps of KEGG pathways and host-pathogen interaction information. Comparative genomes will also be integrated into SpiroESTdb to facilitate the production of large scale synteny maps and their associated genomic information. This database is available on the Web for all users upon registration and provides a large amount of information on potential industrial drug targets and vaccine candidates that can be used in virtual screening initiatives and molecular docking.

Availability

SpiroESTdb is open access and freely available. The curation service requires free user registration because each user needs a unique session. All questions, comments and requests should be sent by email to todaewon@gmail.com.

Project name: SpiroESTdb

Project home page: http://pathod.cdc.go.kr/spiroestdb

Operating system: Linux

Programming languages: HTML, JSP, CSS3, JavaScript, AJAX, Oracle

Other requirements: None

Funding

This work was supported by grant 2009-0084206 from the Ministry of Education, Science and Technology, Korea and grant 2006-N54002-00 from the Korean National Institute of Health. The contribution of the staff at the Genome Research Center of KRIBB in Korea is gratefully acknowledged.

Acknowledgements

The authors thank the pathogen research community at the Korea National Institutes of Health for valuable input on this project and suggestions for building and maintaining this database.

Author details

[1]Division of Malaria and Parasitic Diseases, Korea National Institute of Health, Osong 363-951, Republic of Korea. [2]Genome Resource Center, Korea Research Institute of Bioscience and Biotechnology (KRIBB), Daejeon 305-806, Republic of Korea. [3]University of Science and Technology (UST), Daejeon 303-333, Republic of Korea. [4]Department of Parasitology, Kyungpook National University School of Medicine, Daegu, 700-422, Republic of Korea. [5]Insilicogen, Inc. #909, Suwon, Gyeonggi-do 441-813, Republic of Korea
†These authors contributed equally to this work.

Authors' contributions

DK designed and implemented the database and website and wrote the manuscript, and DK, WY, SN, JP and KL developed the web interfaces, assisted with the design of the database, performed database system administration, and integrated the bioinformatics tools in the application. HY, SC and ML helped with the preparation of the EST data sets (sample collection, cDNA library construction and sequencing). HP and JJ served as the principal investigators of the project. All authors contributed to the writing of the manuscript and have read and approved the final submitted version.

Competing interests

The authors declare that they have no competing interests.

Received: 1 December 2011 Accepted: 8 March 2012
Published: 8 March 2012

References

1. Wiwanitkit V: **A review of human sparganosis in Thailand.** *Int J Infect Dis* 2005, **9**:312-316.
2. Norman SH, Kreutner A Jr: **Sparganosis: clinical and pathologic observations in ten cases.** *South Med J* 1980, **73**:297-300.
3. Park JH, Chai JW, Cho N, Paek NS, Guk SM, Shin EH, Chai JY: **A surgically confirmed case of breast sparganosis showing characteristic mammography and ultrasonography findings.** *Korean J Parasitol* 2006, **44**:151-156.
4. Yun SJ, Park MS, Jeon HK, Kim YJ, Kim WJ, Lee SC: **A case of vesical and scrotal sparganosis presenting as a scrotal mass.** *Korean J Parasitol* 2010, **48**:57-59.
5. Watanabe J, Wakaguri H, Sasaki M, Suzuki Y, Sugano S: **Comparasite: a database for comparative study of transcriptomes of parasites defined by full-length cDNAs.** *Nucleic Acids Res* 2007, **35**:D431-438.
6. Liu F, Chen P, Cui SJ, Wang ZQ, Han ZG: **SjTPdb: integrated transcriptome and proteome database and analysis platform for Schistosoma japonicum.** *BMC Genomics* 2008, **9**:304.
7. Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS, *et al*: **PlasmoDB: a functional genomic database for malaria parasites.** *Nucleic Acids Res* 2009, **37**:D539-543.
8. Aurrecoechea C, Brestelli J, Brunk BP, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS, Heiges M, *et al*: **EuPathDB: a portal to eukaryotic pathogen databases.** *Nucleic Acids Res* 2010, **38**:D415-419.
9. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using phred. I. Accuracy assessment.** *Genome Res* 1998, **8**:175-185.
10. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Repbase Update, a database of eukaryotic repetitive elements.** *Cytogenet Genome Res* 2005, **110**:462-467.
11. Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, *et al*: **TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets.** *Bioinformatics* 2003, **19**:651-652.
12. Huang X, Madan A: **CAP3: A DNA sequence assembly program.** *Genome Res* 1999, **9**:868-877.
13. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403-410.
14. Pruitt KD, Tatusova T, Maglott DR: **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2005, **33**:D501-504.
15. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Res* 1999, **27**:29-34.
16. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, *et al*: **UniProt: the Universal Protein knowledgebase.** *Nucleic Acids Res* 2004, **32**:D115-119.
17. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, *et al*: **InterPro: the integrative protein signature database.** *Nucleic Acids Res* 2009, **37**:D211-215.
18. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**:25-29.
19. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27**:573-580.
20. Barker G, Batley J, H OS, Edwards KJ, Edwards D: **Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP.** *Bioinformatics* 2003, **19**:421-422.
21. Pankov R, Yamada KM: **Fibronectin at a glance.** *J Cell Sci* 2002, **115**:3861-3863.
22. Rostagno AA, Frangione B, Gold LI: **Biochemical characterization of the fibronectin binding sites for IgG.** *J Immunol* 1989, **143**:3277-3282.
23. Leiss M, Beckmann K, Giros A, Costell M, Fassler R: **The role of integrin binding sites in fibronectin matrix assembly in vivo.** *Curr Opin Cell Biol* 2008, **20**:502-507.

doi:10.1186/1756-0500-5-130
Cite this article as: Kim *et al.*: SpiroESTdb: a transcriptome database and online tool for sparganum expressed sequences tags. *BMC Research Notes* 2012 **5**:130.