**Supplementary Information SI1**

Helminth.net: expansions to Nematode.net and an introduction to Trematode.net

John Martin1\*, Bruce A Rosa1\*, Philip Ozersky1, Kymberlie Hallsworth-Pepin1, Xu Zhang1, Veena Bhonagiri-Palsikar1, Rahul Tyagi1, Qi Wang1, Young-Jun Choi1, Xin Gao1, Samantha N. McNulty1, Paul J. Brindley2 and Makedonka Mitreva1, 3, †

1 The Genome Institute, Washington University School of Medicine, St. Louis, MO 63108.

2 Department of Microbiology, Immunology & Tropical Medicine, and Research Center for Neglected Diseases of Poverty, School of Medicine & Health Sciences, The George Washington University, Washington, DC 20037, USA

3 Department of Internal Medicine and Department of Genetics, Washington University School of Medicine, St. Louis, MO 63108.

**Index**

[Trematode.net Data-Mining Portal Usage 3](#_Toc399157176)

[TremaGene 3](#_Toc399157177)

[URL: 3](#_Toc399157178)

[Overview: 3](#_Toc399157179)

[Search Interface: 3](#_Toc399157180)

[Supplementary Figure S1 3](#_Toc399157181)

[Results Page: 4](#_Toc399157182)

[Supplementary Figure S2 4](#_Toc399157183)

[Detailed Report Page: 5](#_Toc399157184)

[TremaBlast 7](#_Toc399157185)

[URL: 7](#_Toc399157186)

[Overview: 7](#_Toc399157187)

[Search Interface: 7](#_Toc399157188)

[Supplementary Figure S4 7](#_Toc399157189)

[BLAST Results : 8](#_Toc399157190)

[Supplementary Figure S5 8](#_Toc399157191)

[BLAST Algorithm: 9](#_Toc399157192)

[TremaBrowse 10](#_Toc399157193)

[URL: 10](#_Toc399157194)

[Overview: 10](#_Toc399157195)

[Gene List Interface: 11](#_Toc399157196)

[Supplementary Figure S6 11](#_Toc399157197)

[GBrowse Results: 12](#_Toc399157198)

[Supplementary Figure S7 12](#_Toc399157199)

[TremaPath 13](#_Toc399157200)

[URL: 13](#_Toc399157201)

[Overview: 13](#_Toc399157202)

[E-value Threshold Interface: 14](#_Toc399157203)

[Supplementary Figure S8 14](#_Toc399157204)

[KEGG Pathway Selection Interface: 15](#_Toc399157205)

[Supplementary Figure S9 15](#_Toc399157206)

[Tremapath Results Interface: 16](#_Toc399157207)

[Supplementary Figure S10 16](#_Toc399157208)

[Microbiome Interaction 17](#_Toc399157209)

[Access: 17](#_Toc399157210)

[Example URL: 17](#_Toc399157211)

[Microbiome Report Example: 17](#_Toc399157212)

[Supplementary Figure S11 17](#_Toc399157213)

[Detailed Methods for Data Production 18](#_Toc399157214)

[SNP calling and annotation 18](#_Toc399157215)

[FPKM calculation 18](#_Toc399157216)

[References 19](#_Toc399157217)

# Trematode.net Data-Mining Portal Usage

## TremaGene

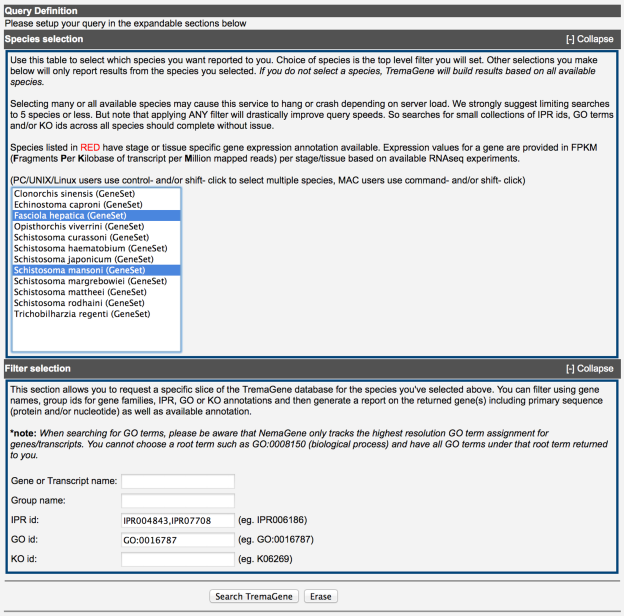
### URL:

<http://www.trematode.net/TN_frontpage.cgi?navbar_selection=tremagene&subnav_selection=tremagene>

### Overview:

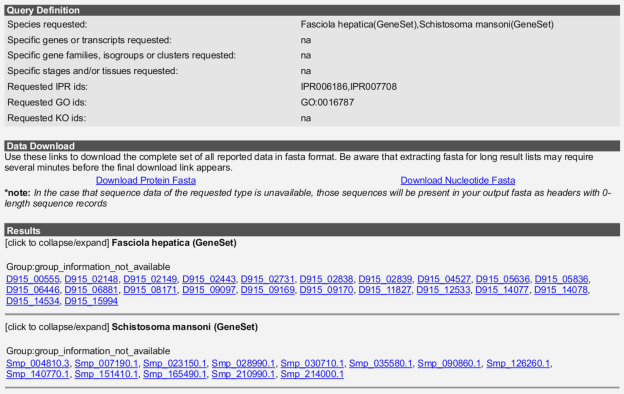
Access into the TremaGene database frequently comes from other tools within the Trematode.net site such as the contig links from TremaPath which directly jump to the details pages that are the terminus of a TremaGene search, or from associated external sites. But the TremaGene Database Search tool can also be used to extract custom slices of our database using available annotations. This tool is also very useful for retrieving the full protein or nucleotide fasta of our genesets, or of the slice you define, using the appropriate Download links.

### Search Interface:

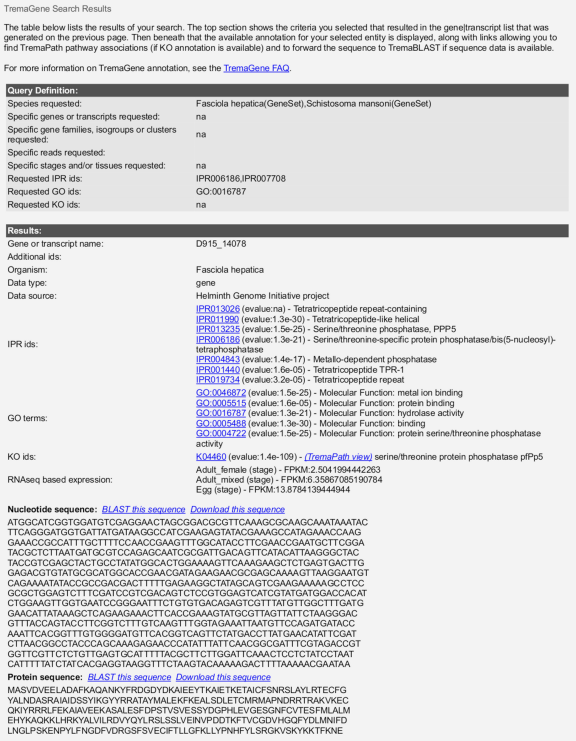


Supplementary Figure S1: The front page of the TremaGene search interface.

### Results Page:



Supplementary Figure S2: The search results page, providing a view of the data subset defined using the submitted filters.

Detailed Report Page: **  
Supplementary Figure S3. The detailed report page. IPR (InterPro), GO, KO and stage specific expression (FPKM) annotation can be seen here, as well as links for downloading protein and nucleotide fasta, for jumping into the TremaPath viewer centering on the reported KO ids and for forwarding the sequence into TremaBlast for further analysis. Additional (but unseen in this example) annotations include status as a putative chokepoint enzyme, and ChEMBL associated drug target information (currently only available in NemaGene).

*Detailed Usage Instructions*

TremaGene can be searched using InterPro, GO and/or KO id filters. First click on the [+] Expand label for the Species selection section and select 1 or more species to start your query. Note that if you select no species from the list, your query will be applied to all species in TremaGene (depending on the complexity of your query, this may take a long time to complete). After selecting the species to focus on, open the sections below to set specific filters you'd like to apply. You're able to request a specific gene name (or comma-delimited list of gene names), orthologous group name if such is available, InterPro id, GO term and/or KO id. You are allowed to enter comma-delimited lists of any of those ids as well. Note that filtering on multiple ids of a single type will return genes/transcripts annotated by any of those ids. But if you set filters using 2 or more id types (i.e. InterPro id + KO id), each gene or transcript returned will be required to have at least one id from each list you supplied.

You will then arrive at a page showing the slice of data you've retrieved from TremaGene. The Query Definition section now displays the query you made to extract the results below. Then the Data Download section allows you to download the full fasta for all the genes/transcripts you requested. Note that if the type of fasta you request (nucleotide or protein) doesn't exist for any of the genes or transcripts in your list, the output file will still display the gene name(s) as headers, but the sequence record for those will be empty. The Results section will list all the genes and/or transcripts in your return set, organized by species, then by group if available (typically groups built by software such as orthomcl or inparanoid). Each gene or transcript name is a link that will take you to a final detail page showing the available annotations for that entity. You will also be able to download that single entity, or forward its sequence into TremaBlast for further analysis. For more information on TremaGene and the available annotations see the TremaGene FAQ.

## TremaBlast

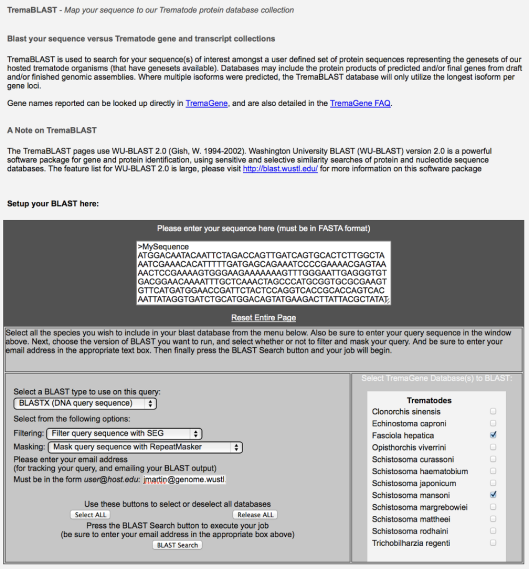
### URL:

<http://www.trematode.net/TN_frontpage.cgi?navbar_selection=tremagene&subnav_selection=tremablast>

### Overview:

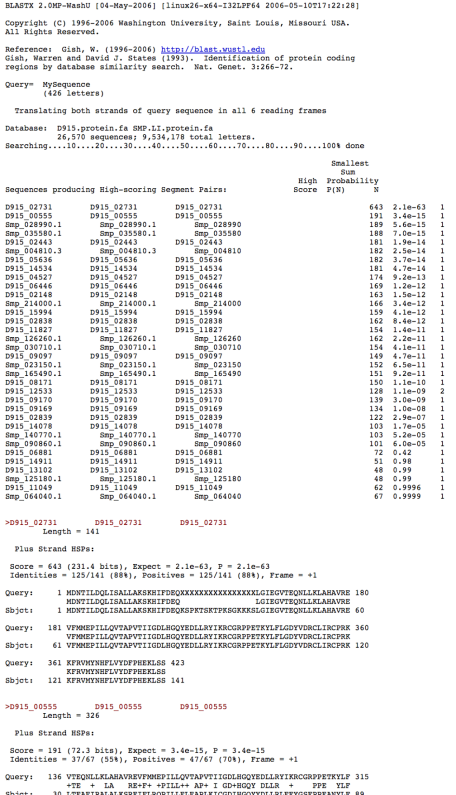
TremaBLAST is used to search for your sequence(s) of interest amongst a user defined set of protein sequences representing the genesets of our hosted trematode organisms (that have genesets available). Databases may include the protein products of predicted and/or final genes from draft and/or finished genomic assemblies. Where multiple isoforms were predicted, the TremaBLAST database will only utilize the longest isoform per gene loci. Gene names reported can be looked up directly in [TremaGene](http://www.trematode.net/NN3_frontpage.cgi?navbar_selection=tremagene&subnav_selection=tremagene), and are also detailed in the [TremaGene FAQ](http://www.trematode.net/NN3_frontpage.cgi?navbar_selection=tremagene&subnav_selection=tremagene_faq).

### Search Interface:



Supplementary Figure S4: The TremaBlast query interface.

### BLAST Results :



Supplementary Figure S5: The TremaBlast results emailed to the user.

### BLAST Algorithm:

The TremaBLAST pages use WU-BLAST 2.0 ([1](#_ENREF_1)). Washington University BLAST (WU-BLAST) version 2.0 is a powerful software package for gene and protein identification, using sensitive and selective similarity searches of protein and nucleotide sequence databases. The feature list for WU-BLAST 2.0 is large, please visit http://blast.wustl.edu/ for more information on this software package.

## TremaBrowse

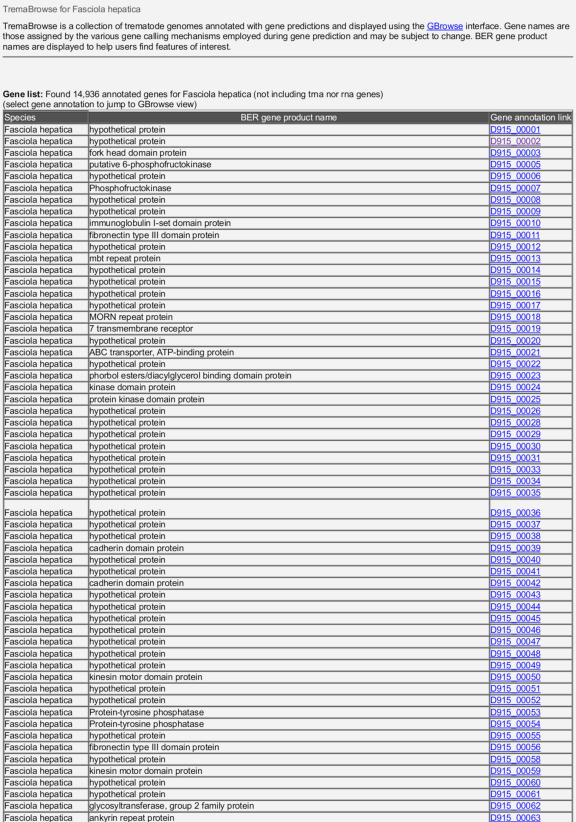
### URL:

<http://www.trematode.net/TN_frontpage.cgi?navbar_selection=tremagene&subnav_selection=tremabrowse>

### Overview:

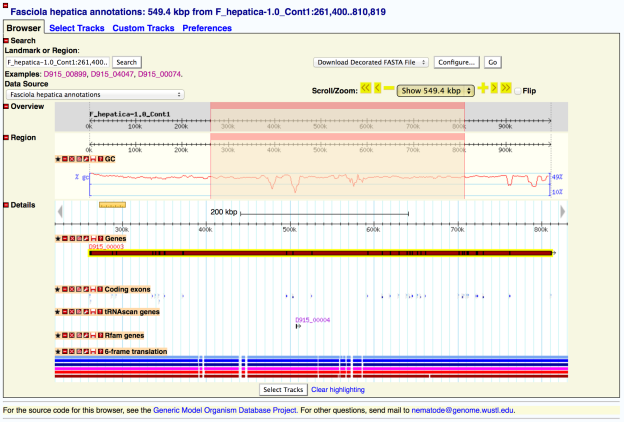
TremaBrowse is a collection of trematode genomes annotated with gene predictions and displayed using the GBrowse ([2](#_ENREF_2)) interface. Gene names are those assigned by the various gene calling mechanisms employed during gene prediction and may be subject to change. Minimal annotation is made available in the gene list using either final gene product information as annotated by the BER pipeline, or information derived from WU-BLAST mappings to NR using a post-alignment cutoff of 35 bits + 55% identity.

### Gene List Interface:

****

Supplementary Figure S6: The TremaBrowse gene list for *Fasciola hepatica*.

### GBrowse Results:



Supplementary Figure S7: The GBrowse interface with annotations, for the gene D915\_00003 from *Fasciola hepatica*.

## TremaPath

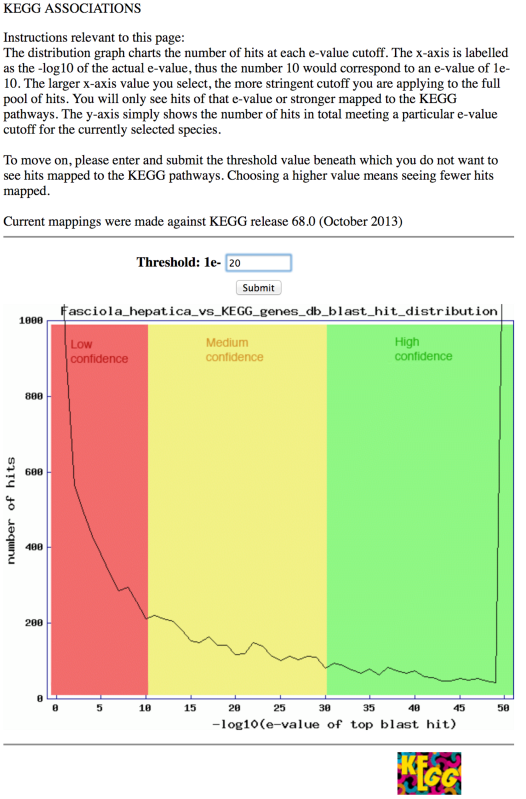
### URL:

<http://www.trematode.net/TN_frontpage.cgi?navbar_selection=comparative_genomics&subnav_selection=species_tremapath>

### Overview:

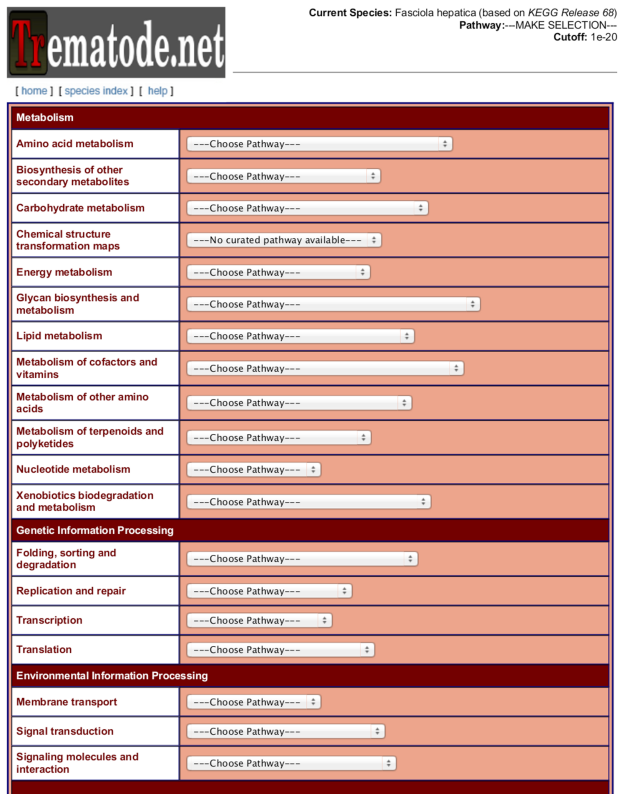
To view KEGG release 68.0 associations for any of the species listed in TremaPath, simply click on “View” in the TremaPath link, then provide an e-value cutoff to recruit alignments to the KEGG defined pathway maps. Mappings are performed using BLAST to align transcripts to the KEGG genes database (the KO annotated only fraction of release 68.0). The KO ids of the subjects hit are used to relate nematode sequences to the KEGG pathway maps. Choosing a less stringent score will be more sensitive, but can introduce false positive mappings.

### E-value Threshold Interface:



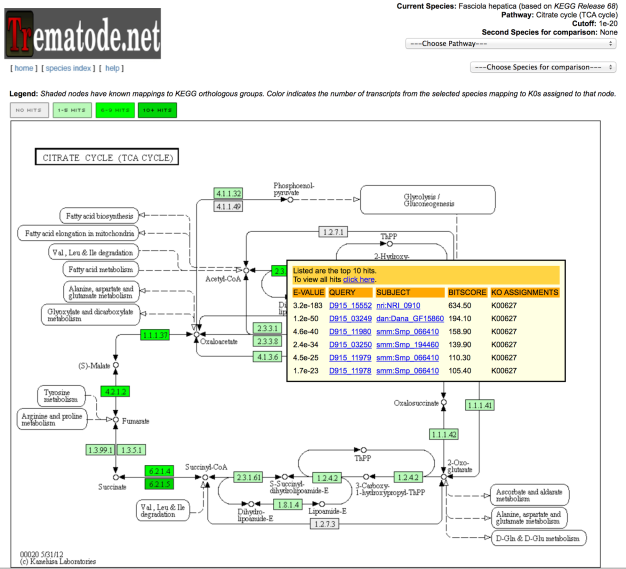
Supplementary Figure S8: The interface for entering an E-value significance cutoff for *Fasciola hepatica*.

### KEGG Pathway Selection Interface:



Supplementary Figure S9: KEGG pathway selection interface.

### Tremapath Results Interface:



Supplementary Figure S10: TremaPath Results for the “Citrate Cycle (TCA Cycle)” pathway in *Fasciola hepatica*.

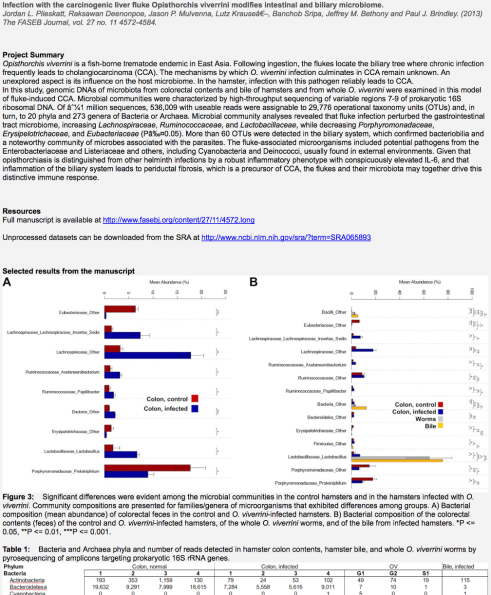
## Microbiome Interaction

### Access:

Microbiome Interaction reports can be found through direct links in the dropdown menu on the main page

Example URL: <http://www.trematode.net/TN_frontpage.cgi?navbar_selection=environmental_interaction&subnav_selection=opisthorchis_viverrini_project_page>

### Microbiome Report Example:



Supplementary Figure S11: An example of a microbiome report on the interaction of *O. viverrini* with its host microbiome.

# Detailed Methods for Data Production

## SNP calling and annotation

NemaBrowse has been updated to feature tracks displaying SnpEff-annotated variants ([3](#_ENREF_3)) from different population of helminthes. Genomics reads from different populations were mapped against the genomic reference hosted in NemaBrowse (an assembly of the susceptible genomic reads themselves) using BWA mem v0.7.5a (http://bio-bwa.sourceforge.net). The alignments were then filtered, re-aligned around indels, and processed using the GATK Unified Genotyper ([4](#_ENREF_4)) to identify snp loci. Post-identification filters were run to remove low confidence polymorphisms, and the results were fed into SnpEff for annotation. Snp loci falling in coding regions were labeled as synonymous or non-synonymous changes (or ‘both’ in cases where more than a single alternate allele was found, and those alternate alleles resulted in different state changes of their amino acid). Annotated loci from the populations are loaded as separate tracks in GBrowse instance hosted in NemaBrowse.

## FPKM calculation

NemaGene and TremaGene now leverage our ever-growing collection of RNAseq data, allowing us to annotate many genes with stage and or tissue-specific normalized expression data in the form of FPKM (Fragments Per Kilobase of transcript per Million mapped reads) values. Available RNAseq for a given species is mapped back to annotated reference using TopHat ([5](#_ENREF_5)), and then htseq-count ([6](#_ENREF_6)) provides a number of reads mapping to each gene. That count is then converted to FPKM for each gene using the formula:

FPKM = count / (gene\_length\_in\_kb \* library\_size)

where the ‘library\_size’ is equal to the sum of reads mapped across all genes in the sample, in millions.

# References

1. Lopez, R., Silventoinen, V., Robinson, S., Kibria, A. and Gish, W. (2003) WU-Blast2 server at the European Bioinformatics Institute. *Nucleic acids research*, **31**, 3795-3798.

2. Stein, L.D. (2013) Using GBrowse 2.0 to visualize and share next-generation sequence data. *Brief Bioinform*, **14**, 162-171.

3. Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X. and Ruden, D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, **6**, 80-92.

4. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*, **43**, 491-498.

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

6. Anders, S., Pyl, P.T. and Huber, W. (2014) *HTSeq: A Python framework to work with high-throughput sequencing data*.