jove Journal of Visualized Experiments www.jove.com

Video Article

IR-TEx: An Open Source Data Integration Tool for Big Data Transcriptomics Designed for the Malaria Vector *Anopheles gambiae*

Victoria A. Ingham[1], Andrew Bennett[2], Duo Peng[3], Simon C. Wagstaff[2], Hilary Ranson[1]

[1]Vector Biology, Liverpool School of Tropical Medicine

[2]Research Computing Unit, Liverpool School of Tropical Medicine

[3]Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health

Correspondence to: Victoria A. Ingham at Victoria.Ingham@lstmed.ac.uk

URL: https://www.jove.com/video/60721
DOI: doi:10.3791/60721

Keywords: Biology, Issue 155, insecticide resistance, transcriptomics, data integration, RNAseq, microarray, detoxification, *Anopheles*, bioinformatics, shiny

Date Published: 1/15/2020

Citation: Ingham, V.A., Bennett, A., Peng, D., Wagstaff, S.C., Ranson, H. IR-TEx: An Open Source Data Integration Tool for Big Data Transcriptomics Designed for the Malaria Vector *Anopheles gambiae*. *J. Vis. Exp.* (155), e60721, doi:10.3791/60721 (2020).

Abstract

IR-TEx is an application written in Shiny (an R package) that allows exploration of the expression of (as well as assigning functions to) transcripts whose expression is associated with insecticide resistance phenotypes in *Anopheles gambiae* mosquitoes. The application can be used online or downloaded and used locally by anyone. The local application can be modified to add new insecticide resistance datasets generated from multiple -omics platforms. This guide demonstrates how to add new datasets and handle missing data. Furthermore, IR-TEx can be completely and easily recoded to use-omics datasets from any experimental data, making it a valuable resource to many researchers. The protocol illustrates the utility of IR-TEx in identifying new insecticide resistance candidates using the the microsomal glutathione transferase, *GSTMS1*, as an example. This transcript is upregulated in multiple pyrethroid resistant populations from Côte D'Ivoire and Burkina Faso. The identification of co-correlated transcripts provides further insight into the putative roles of this gene.

Video Link

The video component of this article can be found at https://www.jove.com/video/60721/

Introduction

The ability to measure the expression of large numbers of transcripts simultaneously through microarray platforms and RNAseq technology has resulted in the generation of vast datasets associating transcript expression with a particular phenotype in both model and non-model organisms. These datasets are an extremely rich resource for researchers, the power of which can be increased by combining relevant sets in a big data integration approach. However, this methodology is limited to those with particular bioinformatics skills. Described here is a program, IR-TEx (previously published by Ingham et al.[1]) that is written in an R package called Shiny[2] and allows users with little bioinformatics training to integrate and interrogate these datasets with relative ease.

IR-TEx, found at http://www.lstmed.ac.uk/projects/IR-TEx, was written to explore transcripts associated with insecticide resistance in *Anopheles gambiae*, the major African malaria vector[1]. Malaria is a parasitic disease caused by *Plasmodium* species, transmitted between humans through the bites of female *Anopheles* mosquitoes. Targeting the mosquito vector with insecticides has proven to be the most effective means of preventing malaria-related morbidity and mortality in Africa. The scaling up of tools (i.e., long lasting insecticidal nets) has also been pivotal in the dramatic reductions in malaria cases since 2000[3]. With a very limited number of insecticides available, there is strong evolutionary pressure on the mosquitoes, and resistance is now widespread in African malaria vectors[4].

Additionally, target site mutations[5] and metabolic clearance of insecticides[6,7] remain the primary studied mechanisms of resistance, but other potent resistant mechanisms are now emerging[1]. Many of these new mechanisms have not previously been associated with insecticide resistance but have been detected by searching for common patterns of gene expression across multiple resistant populations using the IR-TEx app and subsequently functionally validated by genomics approaches[1].

Described here is a step-by-step approach to using IR-TEx, both on the web and when installed locally. The protocol describes how new insecticide resistance datasets can be integrated into the existing package and explains how to operate with missing data. Finally, it describes how to use this software with other -omics datasets that are unrelated to insecticide resistance, thus combining data from varying -omics approaches while also operating with missing values and normalization so that data are comparable.

Protocol

1. Using the IR-TEx web application

1. Running the application in a web browser
    1. Open the IR-TEx web application by following the link at the bottom of the page found at http://www.lstmed.ac.uk/projects/IR-TEx.
    2. Once the web page has initialized, click the **Application** button at the top of the page, which will display the application and associated outputs.
    3. Read each output related to the default entry of **AGAP008212-RA** (*CYP6M2*) in the transcript ID box with the following conditions: *An. coluzzii* datasets that are (i) exposed to pyrethroid insecticides or (ii) not exposed to any insecticide class, and associated transcripts with a correlation of |r| >0.98.

2. Exploring expression of a transcript of interest
    1. To select a transcript of interest, input the transcript ID into the **Transcript ID** box, remembering that transcripts end in **-RX** dependent upon isoform of interest.
    2. Select the datasets to interrogate by ticking the relevant boxes for (i) Countries; (ii) Exposure status, (iii) Species of interest; and (iv) Insecticide class of interest, all while ensuring that these criteria result in >1 included dataset (see Supplementary Table 1 in Ingham et al.[1]).
    NOTE: (iii) refers to the member of the *An. gambiae* species complex that the user interested in. Currently, data are available for *An. coluzzii* and *An. arabiensis*.
    3. Click **Update View** at the bottom of the selection menu or press **Return**, ignoring **Absolute Correlation Value** (for now).
    4. Give the application time to update.
    5. Read the first graph as: $\log_2$ fold change between a resistant population and lab-susceptible mosquito population of the transcript of interest across each dataset that meets the criteria selected in step 1.2 (**Figure 1**). The details of all datasets can be found in Ingham et al.[1].
    6. Read the information below the graph as: the fold changes between the resistant and susceptible mosquitoes for each relevant dataset, in addition to the corrected p-values (Q). Each row represents individual probes on the microarray. The methodology for graphical display has been reported previously[1].
    7. Read the additional table below as the number of experiments in which the transcript of interest is significant as well as the total number of experiments matching the criteria selected in step 1.2.
    8. To download the data in tab separated format, click the **Download** button under the two tables. This allows the user to explore data in an easier manner using a program such as Excel.
    9. Interpret the map as follows: each point represents the approximate collection sites of resistant mosquitoes in each dataset in which the transcript of interest is differentially expressed. The colors follow a traffic light system that is explained in the app (**Figure 2**).
    10. For steps 1.2.5 and 1.2.8, save the graphical outputs by right-clicking, clicking **Save image as…,** and choosing an appropriate folder.
    NOTE: In the instance of an output error by the application, it is likely that no datasets match the inputted criteria. Check Supplementary Table 1 in Ingham et al.[1] if this occurs.

3. Identifying putative functions/pathways of transcript of interest
    1. Correlations (minimum r^2 value inputted) of the expression patterns of transcripts across multiple datasets can be used to predict transcript function and potentially elucidate coregulated transcripts from the same pathway. Using the example from Ingham et al.[1] (AGAP001076-RA; *CYP4G16*), follow steps 1.2.1–1.2.2 in the section above, selecting all datasets for maximal power.
    2. Before clicking **Update View**, move the **Absolute Correlation Value** slider to 0.85, and click **Update View** or press **Return**.
    3. Examine the correlation table (bottommost table) to find the multiple transcripts that are now displayed and are correlated (|r| = 0.85) with the inputted transcript.
    4. Manipulate the **Absolute Correlation Value** slider and observe any changes in the bottommost graph and table; the outputs from step 1.3.2 will remain unchanged. As shown in **Figure 3** (|r| > 0.9, |r| > 0.8), lowering the stringency of the correlation value will show more transcripts but will introduce more noise.
    5. Read the table below the graphical output, which (in addition to the parameters described in step 1.2.6) contains the correlation value for each transcript.
    6. To download the data in a tab-separated format, click the **Download** button.
    7. Functional enrichment analysis can be performed on the downloaded transcript ID list using DAVID analysis[8]. Once on the DAVID website (found at https://david.ncifcrf.gov/), select **Functional Analysis**. Paste the full gene list, using gene IDs [identifier without the -RX, which can be done in excel by inserting a column to the right of the Systematic ID and typing **=LEFT(X1,10)**, where X1 is the Systematic ID cell]. Select the identifier as **VectorBase_ID** and gene list and click **Submit List**.
    8. Click the **Functional Annotation Clustering** button to yield an overview of the enrichments found in this correlation network, allowing a potential function to be assigned to the transcript. Explore in-depth enrichments by looking through the different categories and clicking the **+** buttons for each and subsequently clicking **Chart**.

2. Downloading and implementing IR-TEx locally

1. Downloading and running IR-TEx
    1. Go to the link found at http://github.com/LSTMScientificComputing/IR-TEx; and click **Clone or download** | **Download Zip**. Direct to a folder of choice and unzip the file in that folder.
    2. Download the latest version of R software for the appropriate operating system from the link found at http://cran.r-project.org/mirrors.html. Install the program.

3. Download and install the latest R Studio software, again for the appropriate operating system from the link found at http://www.rstudio.com/products/rstudio/download/.
4. Once installed, open **R Studio** | **Supplemental coding File 1** and run each line to set up the system for IR-TEx.
5. Once all packages are successfully installed and updated as required, go to **File | Open**, locate **IR-TEx.R**, highlight, and open. This should now be visible in the top window of **R Studio**.
6. To run the app, press the **Run App** button in the top-right of the window, and a second window will pop up in which the app will load. Once the loading is complete, for full functionality click **Open in Browser** located in the top-right of the loaded window.

2. Adding resistance datasets to IR-TEx (generated using *Anopheles gambiae* 15k Agilent array)
   1. To add a new analyzed dataset generated on the same microarray platform (A-MEXP-2196) to the available dataset, download the app and locate the unzipped folder downloaded in section 2.1.
   2. Open **Additional File 1**, which represents an output from a limma analysis on A-MEXP-2196 [1]. Using Excel, in column H1, write **Fold_Change**, and in H2, write **=2^B2**, in which B2 is the log fold change. Apply this throughout column H to produce raw fold changes.
   3. Arrange **Additional File 1** such that column A is the ID, column B is the fold change from column H (copy column H, highlight column B, then right-click and paste values) and column C is the adjusted p-value. Delete all other columns and save as a tab-delimited file.
   4. Open **Supplemental coding File 2** and run using the tab-delimited sheet produced in step 2.2.3.
      NEWFILE_FC = c('COUNTRY','EXPOSURE STATUS','SPECIES','INSECTICIDE')
      NEWFILE_Q = c('COUNTRY','EXPOSURE STATUS','SPECIES','INSECTICIDE')
      NOTE: Fields within single quotation marks should be changed to reflect information from the new dataset. Exposure status refers to whether samples were collected following insecticide exposure (exposed/unexposed). Insecticide: if 'unexposed', use 'none'. See Fold_Changes.txt. for metadata from other samples. Ensure that spelling is consistent.
   5. Open **geography.txt**, scroll to the final occupied row, and select below. Type in the name of the dataset, followed by **Q** and **NEWFILE_Q** in column 1, the latitude of the sample collection site in column 2, and the longitude in column 3. Save the changes.
   6. If any novel entries are used (i.e., Gambia), which are not available for selection in the dataset (see Ingham et al. Supplementary Table 1[1]), these will need to be added into the code. To do this, open IR-TEx.R in RStudio and locate line 26 as indicated by RStudio, at which point the following should begin:
      '**sidebarPanel(….'.**
      NOTE: Each of the proceeding rows relates to an item of metadata inputted into the rows below the dataset name in Fold_Changes.txt in step 2.2.5.
   7. To add the novel metadata, scroll to the end of the line of the metadata of choice, and locate the term 'selected='. Immediately following this should be a comma and closed bracket; at this point, click the cursor within the closed bracket. After the final apostrophe, type a comma, followed by an apostrophe, followed by the new metadata (e.g., 'Gambia'), and save the changes. See below for an example.
      checkboxGroupInput('CountryInput','Select Relevant Countries',c('Burkina Faso','Cote D`Ivoire','Cameroon','Equatorial Guinea','Zambia','Tanzania','Sudan','Uganda','Togo', **'Gambia'**),selected=c('Burkina Faso','Cote D`Ivoire','Cameroon','Equatorial Guinea','Zambia','Tanzania','Sudan','Uganda','Togo'))
   8. Run the app. The new metadata entry should appear as an unselected tick box under the relevant heading. If the user wants it to be selected, it should be added after the selected=c(…, as shown below:
      *checkboxGroupInput('CountryInput','Select Relevant Countries',c('Burkina Faso','Cote D`Ivoire','Cameroon','Equatorial Guinea','Zambia','Tanzania','Sudan','Uganda','Togo', **'Gambia'**),selected=c('Burkina Faso','Cote D`Ivoire','Cameroon','Equatorial Guinea','Zambia','Tanzania','Sudan','Uganda','Togo', **'Gambia'**))*
   9. To add resistance datasets not performed on A-MEXP-2196, see section 3.

3. Modifying IR-TEx for use with different datasets

1. Use across multiple -omics platforms and proceeding with missing data
   1. To proceed with "0" in datasets: consult the dataset source for the specific meaning of "0". It is recommended that "0" is (conservatively) replaced with "NA". As with raw fold changes (B/A), "0" indicates an undetected signal in experimental condition B. In the case that experimental condition A exhibits substantial expression, the user can apply a small fold change value.
   2. Open **Additional File 2.txt**, an RNAseq file adapted from Uyhelji et al.[9]. This file represents the template in which new data should be based: column A = identifier, column B = raw fold change, and column C = adjusted p-value. Use this file to run through the below steps.
   3. Run the R code to match identifiers into a single tab-delimited file across platforms, then organize and normalize the data (**Supplemental Coding File 2**). Instructions are contained within the file. Any FILEPATH will be separated by "/" for MacOS or "//" for Windows (change these from "\", as they will appear).
   4. Output the file produced at the end of **Supplemental Coding File 2** to a location of choice for use in step 3.1.5. **Supplemental Coding File 2** will output a new **Fold_Changes.txt** file. Back up the original file.
   5. Execute the code contained in **Supplementary coding File 3**. Find the output file named **FC_distribPlot.png** in the folder specified as **FILEPATH**. Check the distributions of $\log_2$ fold change to verify that the $\log_2$ fold change distributions are nearly identical across datasets.
   6. Follow instructions from step 2.2.6 to edit additional files and ensure compatibility of the new **Fold_Changes.txt**.

2. Modifying IR-TEx for use with completely new datasets
   1. Open **IR-TEx.R** in RStudio and locate the lines (23–34) beginning with:
      '***tabPanel('***
      and ending in:
      submitButton("Update View", icon("refresh"))
      ),

2. Change the **AGAP008212-RA** found in the below lines to a transcript of interest in the new data.
   textInput('textInput','Transcript ID',value='AGAP008212-RA'),
3. Locate the four options beginning with:
   checkboxGroupInput(
   These options can be modified to represent important metadata that the user wishes to filter the new data by. In each instance, the user should change the **Select Relevant Countries**; **Select Exposure Status**; **Select Relevant Species;** and **Select Insecticide Class** to be representative of the data (i.e., **Select Tissue Type**; **Select Sex**; **Select Age Bracket**; **Select Disease Status**).
4. Identify the metadata associated with the dataset and input to replace the existing options immediately after the first ***c('***. In each instance, the options will be contained within speech marks and separated from the next selection by a comma. After the final selection, the bracket should be closed. An example for **Select Disease Status** is:
   c('Infected', 'Uninfected', 'Unknown')
5. Choose which of these metadata will be selected upon opening the app. These can be changed by modifying the options after ***selected=c('***. An example for **Select Disease Status** is:
   selected=c('Infected', 'Uninfected')
   This will instruct the app to select only datasets matching these criteria on initial loading.
6. To create a new data table, follow the layout found in **Fold_Changes.txt** and instructions in section 2. Change the metadata to each respective change outlined in step 3.2.4, exactly as written into the code (R is case-sensitive). Into the detoxification column, input gene names, and in the transcript type column, input gene descriptions for each transcript. Follow section 3.2 when adding new datasets.
7. If mapping is not relevant to the experimental requirements, locate the following lines of code and place '#' in front:
   Lines 49–51:
   br(),br(),
   withSpinner(plotOutput("Geography")),
   textOutput('Geography_legend'),
   Lines 493 starting:
   output$Geography <- renderPlot({
   To line 602 ending:
   output$Geography_legend <- renderText({
   paste("Significant Transcripts Only (p", as.expression("<="),"0.05): FC > 5 = Red, FC > 1 = Amber, FC < 1 = Green",sep="")
   })

Representative Results

Using the Fold_Changes.txt file included with IR-TEx, we compared transcripts that were significantly differentially expressed in resistant *Anopheles coluzzii* and *Anopheles gambiae* datasets to susceptible controls from Côte D'Ivoire and Burkina Faso. This yielded 18 transcripts of interest (**Table 1**; this search can be performed using Excel, R, or other programs). Two of these, an ATPase (AGAP006879) and α-crystallin (AGAP007160), have been previously reported, with the former having a significant effect on pyrethroid resistance[1]. In addition to these two transcripts, two detoxification transcripts, *GSTMS1* (FC_{μ} = 1.95 and 1.85) and *UGT306A2* (FC_{μ} = 2.29 and 2.28) were present.

qPCR validation of two of these transcripts (*GSTMS1*, a detoxification transcript; and AGAP009110-RA, an unknown, mosquito-specific transcript containing a β-1,3-glucan binding domain) were performed as previously described[1]. Analysis was performed using primer sets described in **Additional File 3** and showed that these transcripts were significantly upregulated in a multiresistant population from Côte D'Ivoire (Tiassalé) and another from Burkina Faso (Banfora), compared to the lab-susceptible N'Gousso (**Figure 4A**).

As both transcripts showed significant upregulation in each of the resistant populations, RNAi-induced knockdown was performed on mosquitoes from the LSTM laboratory Tiassalé colony. This colony originates from Côte D'Ivoire and is resistant to all major classes of insecticide used in public health, as previously described[1,10]. Attenuation of expression of *GSTMS1* resulted in a significant increase ($p = 0.021$) in mortality after deltamethrin exposure compared to GFP-injected controls, demonstrating the importance of this transcript in pyrethroid resistance (**Figure 4B**). Conversely, AGAP009110-RA knockdown resulted in no significant ($p = 0.082$) change in mortality after exposure (**Figure 4B**).

GSTMS1 is a microsomal GST and is one of three found in *A. gambiae* mosquitoes[11]. Although members of the epsilon and delta classes of GSTs have been previously implicated in insecticide detoxification[12,13,14], this is the first evidence to our knowledge for a role of microsomal GSTs in pyrethroid resistance[15]. To explore the putative function of this transcript in *Anopheles gambiae sl* mosquitoes, the expression and correlation in IR-TEx were identified. *GSTMS1* was significantly overexpressed in 20 out of 21 datasets available for these species, with the exception of Bioko Island. In each location, the overexpression was less than five-fold compared to the susceptible populations (**Figure 5**).

As microsomal GSTs have largely been ignored as potential insecticide detoxifiers, little is known about their role in insecticide resistance[15]. By exploring the co-correlation of other transcripts, putative functions can be elucidated through the assumption of coregulation or involvement in the same pathways. To maximize power in the correlation network, all microarray datasets present in IR-TEx were selected, and an $|r|$ of >0.75 was selected. **Table 2** shows the output from IR-TEx.

These transcripts are enriched in oxioreductase activity and glucose/carbohydrate metabolism in DAVID's functional annotation tool[8]. Both glucose-6-phosphate dehydrogenase and cytathione gamma-lyase maintain the level of glutathione in mammalian cells[16,17] and thus link directly with *GSTMS1*, a glutathione-S-transferase. Catalase is a fast-acting oxidative stress responder that protects cells from reactive oxygen species damage, a byproduct of pyrethroid exposure. Valacyclovir hydrolase is a hydrolase that may play a role in detoxification in mammalian cells[18]. *CYP4H17* is also present in the correlation network. Cytochrome p450s are direct metabolizers of pyrethroid insecticides, and these breakdown products can be further metabolized by GSTs. Finally, *CYP4H17* has been implicated in pyrethroid resistance in *A. funestus*[19]. Taken together, these data strongly support a role for *GSTMS1* in xenobiotic detoxification.

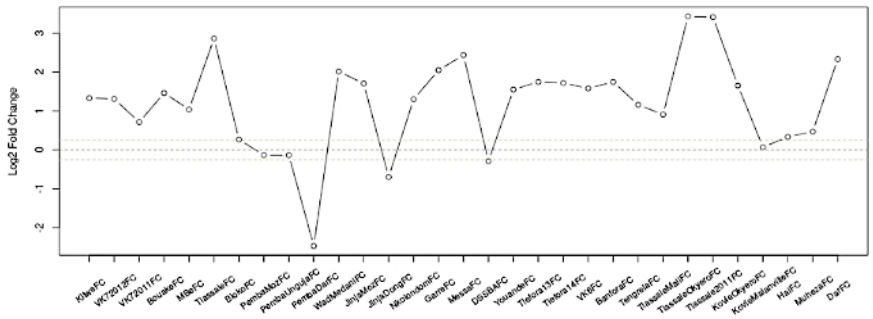

Figure 1: Log$_2$ fold change of AGAP002865-RA in all datasets. The x-axis details the different datasets, information for which can be found in Supplementary Table 1 in a previous publication[1], and the y-axis shows the log$_2$ fold change in the transcript of interest. The light-grey dotted lines indicate approximate thresholds for significance, taken here to be a fold change of <0.8 or fold change of >1.2. The dotted black line indicates a fold change of 1 (i.e., no difference in expression between the resistant and susceptible populations). Please click here to view a larger version of this figure.

Figure 2: Distribution of microarrays showing significant differential expression of AGAP002865-RA in resistant populations. Fold changes are represented in a traffic light system: green fold change of <1, orange fold change of >1, and red fold change of >5. Only datasets with significant ($p \leq 0.05$) differential expression are shown. Please click here to view a larger version of this figure.

Figure 3: Correlation networks of AGAP001076-RA (*CYP4G16*). Pairwise correlations are calculated across all transcripts across the 31 microarray datasets, with a user-defined cut-off applied. Shown here is (**A**) |r| > 0.9 and (**B**) |r| > 0.8. All transcripts displayed on the graph meet this criterion and follow the expression changes of AGAP001076-RA. Please click here to view a larger version of this figure.

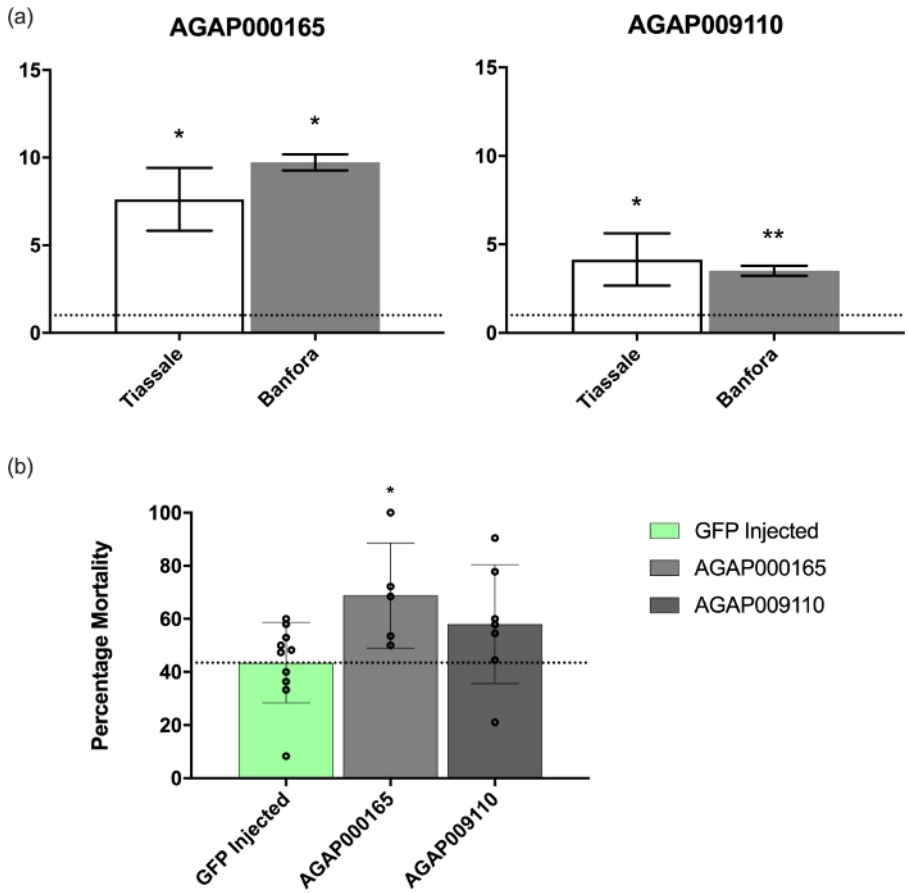

Figure 4: mRNA expression and phenotype upon attenuation of *GSTMS1* and AGAP009110-RA. (**A**) mRNA expression of GSTMS1 and AGAP009110-RA in two multi-resistant *An. coluzzii* populations from Côte D'Ivoire and Burkina Faso, respectively. Levels were compared to the lab-susceptible *An. coluzzii* N'Gousso. Significance levels calculated by ANOVA with a post-hoc Dunnett's test. (**B**) RNAi-induced attenuation of both transcripts compared to GFP-injected controls. *GSTMS1* attenuation shows significant increase in mortality after deltamethrin exposure (calculated by ANOVA with a post-hoc Tukey test; *p ≤ 0.05, **p ≤ 0.01). Please click here to view a larger version of this figure.

Figure 5: Expression of *GSTMS1* in *Anopheles gambiae* and *Anopheles coluzzii* populations. Map showing the significantly differential expression of *GSTMS1* in available microarray datasets. *GSTMS1* was found to be significantly differential in 20 out of 21 microarray datasets. Please click here to view a larger version of this figure.

Transcript ID	Description	Burkina Faso	Côte D'Ivoire
AGAP006879-RA	ATPase	27.94	43.05
AGAP007160-RB	a-crystallin	11.49	10.58
AGAP007160-RC	a-crystallin	11.14	10.38
AGAP007160-RA	a-crystallin	9.78	9.84
AGAP009110-RA	Unknown	9.26	5.96
AGAP007780-RA	NADH dehydrogenase	10.49	3.77
AGAP006383-RA	oligosaccharyltransferase complex subunit beta	3.69	5.57
AGAP007249-RB	Flightin	4.61	3.86
AGAP003357-RA	RAG1-activating protein 1-like protein	4.31	4.05
AGAP007249-RA	Flightin	4.48	3.46
AGAP001998-RA	mRpS10	3.46	2.85
AGAP007589-RA	UGT306A2	2.29	2.28
AGAP000165-RA	GSTMS1	1.95	1.85
AGAP002101-RA	isoleucyl-tRNA synthetase	0.57	0.59
AGAP002969-RA	asparaginyl-tRNA synthetase	0.45	0.45
AGAP004199-RA	solute carrier family 5 (sodium-coupled monocarboxylate transporter), member 8	0.35	0.48
AGAP004684-RA	rRNA-processing protein CGR1	0.36	0.22
AGAP006414-RA	Cht8	0.024	0.36

Table 1: Transcripts significantly differential in the same fold change direction across Burkina Faso and Côte D'Ivoire populations. Transcript ID, gene description, and average fold change for each dataset from the two countries representing *An. coluzzii* and *An. gambiae* populations.

Correlation	Systematic Name	Transcript Type
1	AGAP000165-RA	GSTMS1
0.82	AGAP004904-RA	Catalase
0.76	AGAP007243-RA	26S protease regulatory subunit 8
0.79	AGAP008358-RA	CYP4H17
0.76	AGAP009436-RA	Valacyclovir hydrolase
0.75	AGAP010739-RA	Glucose-6-phosphate 1-dehydrogenase
0.85	AGAP011172-RA	cystathionine gamma-lyase
0.76	AGAP012678-RA	Glucose-6-phosphate 1-dehydrogenase

Table 2: Transcripts co-correlated with *GSTMS1*. The table shows output of the correlation network for *GSTMS1* on IR-TEx with |r| of >0.75. The table shows the Spearman's correlation, transcript ID, and gene description for each co-correlated transcript.

Additional File 1: Output file from A-MEXP-2196 array analyzed on limma. The file originates from a *Met* knockdown compared to a *GFP* control array, described in more detail in ArrayExpress (E-MTAB-4043) and another previous publication[1]. Columns represent AGAP identifier (SystematicName), log fold change (logFC), log expression values (AveExpr), t-statistic (t), uncorrected p-value (P.Value), adjusted p-value (adj.P.Val), and B statistic (B)[20]. For the purposes of this file, the mosquitoes are *Anopheles coluzzi* from Côte D'Ivoire and are unexposed to insecticides, with a collection latitude and longitude of -5.4 and 6.0, respectively. Please click here to view this file (Right click to download).

Additional File 2: Output file from RNAseq experiment. RNAseq analysis taken from Uyhelji et al.[9] describing changes in the transcriptome of *Anopheles* mosquitoes when exposed to 50% salinity. This file is adapted from Table S2 of the publication and includes AGAP identifier (SystematicID), raw fold change (Fold_Change), and adjusted p-value (q_value). Please click here to view this file (Right click to download).

Additional File 3: Primer list for representative results. AGAP identifier, gene name, dsRNA forward, dsRNA reverse, qPCR forward, and qPCR reverse primer sets for each transcript. Please click here to view this file (Right click to download).

Supplemental coding File 1. Please click here to view this file (Right click to download).

Supplemental coding File 2. Please click here to view this file (Right click to download).

Supplemental coding File 3. Please click here to view this file (Right click to download).

Discussion

Big data transcriptomics produces lists of thousands of transcripts that are differentially expressed for each experimental condition. Many of these experiments are performed on related organisms and phenotypes and are almost exclusively analyzed as independent experiments. Utilizing these rich data sources by examining the data holistically and without theoretical assumptions will 1) lead to the identification of new candidate transcripts and 2) prevent the discarding of valuable data simply because there is too much information to validate in vivo[1].

IR-TEx provides users with a limited bioinformatics background with the ability to easily examine multiple datasets, visualize changes in the datasets, and download the associated information[1]. Although IR-TEx does not support searching for more than one transcript in each search, users can examine the associated Fold_Changes.txt files simply by using Excel, R, or other appropriate programs. Further utility of IR-TEx stems from the use of correlation networks to predict transcript function, input of hypothetical proteins or transcripts with unknown functions and use of downstream software to search for enrichments[1].

In the example demonstrated in this protocol, IR-TEx is used according to its original function. Here, it allows exploration of transcripts associated with insecticide resistance and visualization of the distribution of over- and under-expression through mapping graphics. Transcripts of interest are validated in vivo to determine whether the over- or under-expression of given transcripts contributes to an observed phenotype[1] (e.g., insecticide resistance). It was demonstrated here, as previously reported[1], that a dataset can be used in a hypothesis-driven approach to identify transcripts of interest on a country-specific basis. IR-TEx can then be used to 1) explore expression of the transcript and 2) contextualize the transcript's function by applying a pairwise correlation network across all transcripts contained in each -omics dataset. Here, *GSTMS1* was shown to be co-correlated with a number of other transcripts implicated in detoxification. This data (along with knockdown of the transcript that resulted in a significant increase in mortality after insecticide exposure) demonstrates the importance of this transcript in xenobiotic clearance.

IR-TEx represents a valuable resource for exploring insecticide resistance-related transcripts on the web or using local applications. This protocol demonstrates how to modify IR-TEx for different -omics platforms as well as completely new data. The guide illustrates how to use IR-TEx to integrate data from multiple -omics platforms and datasets with missing data as well as how to recode IR-TEx simply so it is useful for anyone researching transcriptomic datasets.

Disclosures

The authors have nothing to disclose.

Acknowledgments

This work was funded by an MRC Skills Development Fellowship to V.I. (MR/R024839/1) and Royal Society Challenge Grant (CH160059) to H.R.

References

1. Ingham, V.A., Wagstaff, S., Ranson, H. Transcriptomic meta-signatures identified in Anopheles gambiae populations reveal previously undetected insecticide resistance mechanisms. *Nature Communications.* **9** (1), 5282 (2018).
2. Chang, W., Cheng, J., Allaire, J., Xie, Y., McPherson, J. *shiny: Web Application Framework for R.* (2017).
3. Bhatt, S. *et al.* The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015. *Nature.* **526** (7572), 207-211 (2015).
4. Ranson, H., Lissenden, N. Insecticide Resistance in African Anopheles Mosquitoes: A Worsening Situation that Needs Urgent Action to Maintain Malaria Control. *Trends in Parasitology.* **32** (3), 187-196 (2016).
5. Donnelly, M. J. et al. Does kdr genotype predict insecticide-resistance phenotype in mosquitoes? *Trends in Parasitology.* **25** (5), 213-219 (2009).
6. Stevenson, B. J. *et al.* Cytochrome P450 6M2 from the malaria vector Anopheles gambiae metabolizes pyrethroids: Sequential metabolism of deltamethrin revealed. *Insect Biochemistry and Molecular Biology.* **41** (7), 492-502 (2011).
7. Müller, P. *et al.* Field-Caught Permethrin-Resistant Anopheles gambiae Overexpress CYP6P3, a P450 That Metabolises Pyrethroids. *PLoS Genetics.* **4** (11), e1000286 (2008).
8. Huang, D. *et al.* The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biology.* **8** (9), R183 (2007).
9. Uyhelji, H. A., Cheng, C., Besansky, N. J. Transcriptomic differences between euryhaline and stenohaline malaria vector sibling species in response to salinity stress. *Molecular Ecology.* **25** (10), 2210-2225 (2016).
10. Edi, C. V., Benjamin, K. G., Jones, C. M., Weetman, D., Ranson, H. Multiple-Insecticide Resistance in Anopheles gambiae Mosquitoes, Southern Côte d'Ivoire. *Emerging Infectious Diseases.* **18** (9), 1508-1511 (2012).
11. Ding, Y., Ortelli, F., Rossiter, L., Hemingway, J., Ranson, H. The Anopheles gambiae glutathione transferase supergene family: annotation, phylogeny and expression profiles. *BMC Genomics.* **4** (1), 1-16 (2003).
12. Enayati, A. A., Ranson, H., Hemingway, J. Insect glutathione transferases and insecticide resistance. *Insect Molecular Biology.* **14** (1), 3-8 (2005).
13. Ranson, H. *et al.* Identification of a novel class of insect glutathione S-transferases involved in resistance to DDT in the malaria vector Anopheles gambiae. *TheBiochemical J*ournal. **359**, 295-304 (2001).
14. Riveron, J. M. *et al.* A single mutation in the GSTe2 gene allows tracking of metabolically based insecticide resistance in a major malaria vector. *Genome Biol*ogy. **15** (2), R27 (2014).
15. Pavlidi, N., Vontas, J., Van Leeuwen, T. The role of glutathione S-transferases (GSTs) in insecticide resistance in crop pests and disease vectors. *Current Opinion in Insect Science.* **27**, 97-102 (2018).
16. Salvemini, F. et al. Enhanced glutathione levels and oxidoresistance mediated by increased glucose-6-phosphate dehydrogenase expression. *Journal of Biological Chemistry.* **274** (5), 2750-2757 (1999).
17. Deplancke, B., Gaskins, H. R. Redox control of the transsulfuration and glutathione biosynthesis pathways. *Current Opinion in Clinical Nutrition & Metabolic Care.* **5** (1), (2002).
18. Puente, X. S., López-Otn, C. Cloning and expression analysis of a novel human serine hydrolase with sequence similarity to prokaryotic enzymes involved in the degradation of aromatic compounds. *Journal of Biological Chemistry.* **270** (21), 12926-12932 (1995).
19. Riveron, J. M. *et al.* Genome-wide transcription and functional analyses reveal heterogeneous molecular mechanisms driving pyrethroids resistance in the major malaria vector Anopheles funestus across Africa. *G3: Genes, Genomes, Genetics.* **7** (6), 1819-1832 (2017).
20. Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology.* **3** (1), 3 (2004).