Supplementary Note 1: IR-TEx User Guide

User Guide

IR-TEx: Insecticide Resistance Transcript Explorer

V.A Ingham, D. Peng, S. Wagstaff and H. Ranson

LSTM
LIVERPOOL SCHOOL
OF TROPICAL MEDICINE

Contents

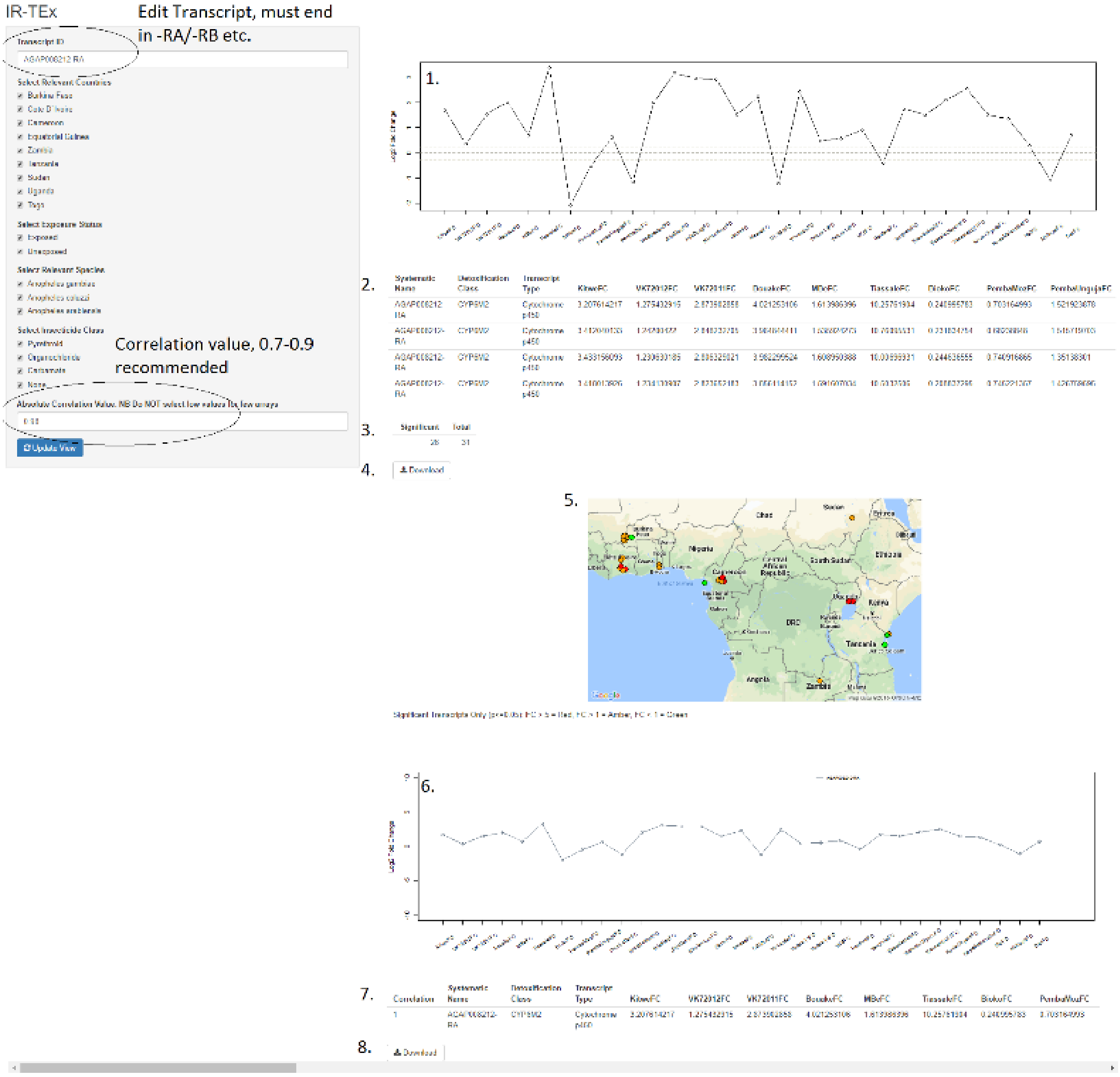
Section 1: Introduction

IR-TEx is an app written in ShinyR to explore microarray datasets that compare resistant and susceptible *Anopheles gambiae*, *An. coluzzi* and *An. arabiensis* populations, available in public repositories, in a used friendly manner. In its current form, IR-TEx allows the user to search for transcripts of interest using a VectorBase Transcript ID by: Country; Exposure Status; Species and Insecticide Class. The user can also find co-correlated transcripts across experiments by manipulating the Absolute Correlation Value (recommended: 0.7-0.9). The outputs from IR-TEx come in several forms.

IR-TEx basics

IR-TEx can be used to explore the relationships between expression levels of transcripts across populations of Anopheline vectors with varying levels of resistance to insecticides. To run the IR-TEx simply visit the GitHub page which will contain a link to the current application web page. IR-TEx is currently hosted at: https://www.lstmed.ac.uk/projects/ir-tex

The output **below** shows the appearance of a typical ***Interactive Dashboard*** displaying transcript expression, experiment and geographical location.



IR-TEx

Edit Transcript, must end in -RA/-RB etc.

Transcript ID

AGAP008212-RA

Select Relevant Countries

☑ Burkina Faso
☑ Cote D'Ivoire
☑ Cameroon
☑ Equatorial Guinea
☑ Zambia
☑ Tanzania
☑ Sudan
☑ Uganda
☑ Togo

Select Exposure Status

☑ Exposed
☑ Unexposed

Select Relevant Species

☑ Anopheles gambiae
☑ Anopheles coluzzi
☑ Anopheles arabiensis

Select Insecticide Class

☑ Pyrethroid
☑ Organochloride
☑ Carbamate
☑ None

Correlation value, 0.7-0.9 recommended

Absolute Correlation Value. NB Do NOT select low values for few arrays

0.98

Update View

1.

2.

Systematic Name	Detoxification Class	Transcript Type	KisumFC	VK72012FC	VK72011FC	DouakoFC	MDoFC	TiassaleFC	DiokoFC	PembaMicoFC	PembaUngujaFC
AGAP008212-RA	CYP6M2	Cytochrome p450	3.207614217	1.275432915	2.873582858	4.021253106	1.613986396	10.25761904	0.240555783	0.703164993	1.521923678
AGAP008212-RA	CYP6M2	Cytochrome p450	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
AGAP008212-RA	CYP6M2	Cytochrome p450	3.433156093	1.230630185	2.886325821	3.582295524	1.608950388	10.00999331	0.244636555	0.740916865	1.35138301
AGAP008212-RA	CYP6M2	Cytochrome p450	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

3.

Significant	Total
28	31

4. Download

5.

Significant Transcripts Only (p<=0.05): FC > 5 = Red, FC > 1 = Amber, FC < 1 = Green

6.

7.

Correlation	Systematic Name	Detoxification Class	Transcript Type	KisumFC	VK72012FC	VK72011FC	DouakoFC	MBoFC	TiassaleFC	DiokoFC	PembaMicoFC
1	AGAP008212-RA	CYP6M2	Cytochrome p450	3.207614217	1.275432915	2.873582858	4.021253106	1.613986396	10.25761904	0.240555783	0.703164993

8. Download

The interactive dashboard is composed of the following:

1) ***Expression Line Graph*** showing the $\log_2$ fold change of the transcript of interest for each microarray data set
2) ***Probe Expression Table*** showing the VectorBase Transcript ID, Detoxification Class, Transcript Description, raw Fold Change (FC) and Q value (Q) (adjusted p-value) for each **probe** (row) and dataset (column).
3) ***Summary Data*** showing the number of arrays in which the transcript is significantly differential.
4) ***Download*** to obtain a local copy of the Probe Expression Table.
5) ***Map*** highlighting the location of the data set containing the transcript of interest with significant differential expression illustrated as a traffic light system.
6) ***Correlation Line Graph*** showing the $\log_2$ fold change of the transcript of interest and transcripts correlated with the transcript of interest (if any) with an absolute correlation greater than that user defined threshold.
7) ***Correlated Transcript Expression Table*** showing the VectorBase Transcript ID, Detoxification Class, Transcript Description, raw Fold Change (FC) and Q value (Q) (adjusted p-value) for each **transcript of interest or correlated transcript** (row) and dataset (column).
8) ***Download*** to obtain a local copy of the Correlated Transcript Expression Table

Performance and Resources

IR-TEx requires only a relatively modest amount of computing power per user. The most computationally intensive part of the application is the initial generation of the correlation matrix. The default matrix that loads on application startup includes an optimum number of datasets and correlation threshold which typically takes ~4s to load on standard Intel i7 processor, consuming ~3GB of RAM in the process. IR-TEx is best ran locally if regular use is intended.

NB. Recalculation of the matrix is required each time a dataset is added or subtracted or an option changed. An increased amount of processing power is required if, for example, all studies are included or a low correlation threshold is selected.

Installing IR-TEx

IR-TEx is a ShinyR application and can be downloaded and installed for free from the following GitHub site at https://github.com/LSTMScientificComputing/IR-TEx and includes a number of files, including a table of fold change and Q values, and a longitude-latitude file for geographical locations of the collection sites. To install locally, the app needs to be installed alongside the packages ggmap (https://cran.r-project.org/web/packages/ggmap/index.html), mapproj (https://cran.r-project.org/web/packages/mapproj/index.html), shinycssloaders (https://cran.r-project.org/web/packages/shinycssloaders/index.html) and ShinyR. Instructions for the latter are available here - https://shiny.rstudio.com.

Section 2: Inputting other resistance datasets

Overview of Entering Data

All datasets used in this app are currently from the LSTM Agilent 15K array V1 (A-MEXP-2196), dating from AgamP3.5 (2009). Although this array is the most commonly used for insecticide resistance experiments due to the multiple probes for 'detoxification family' genes, we recognise that other array designs may be used in the future. There are two options for inputting these datasets; the first is to use fold changes and Q values for probes only found on the original arrays and the second is to set missing probes to '0' which would cause them to be missing within the app. Below is a walkthrough to adding more resistance datasets to the existing data, without having to change any core code within the app.

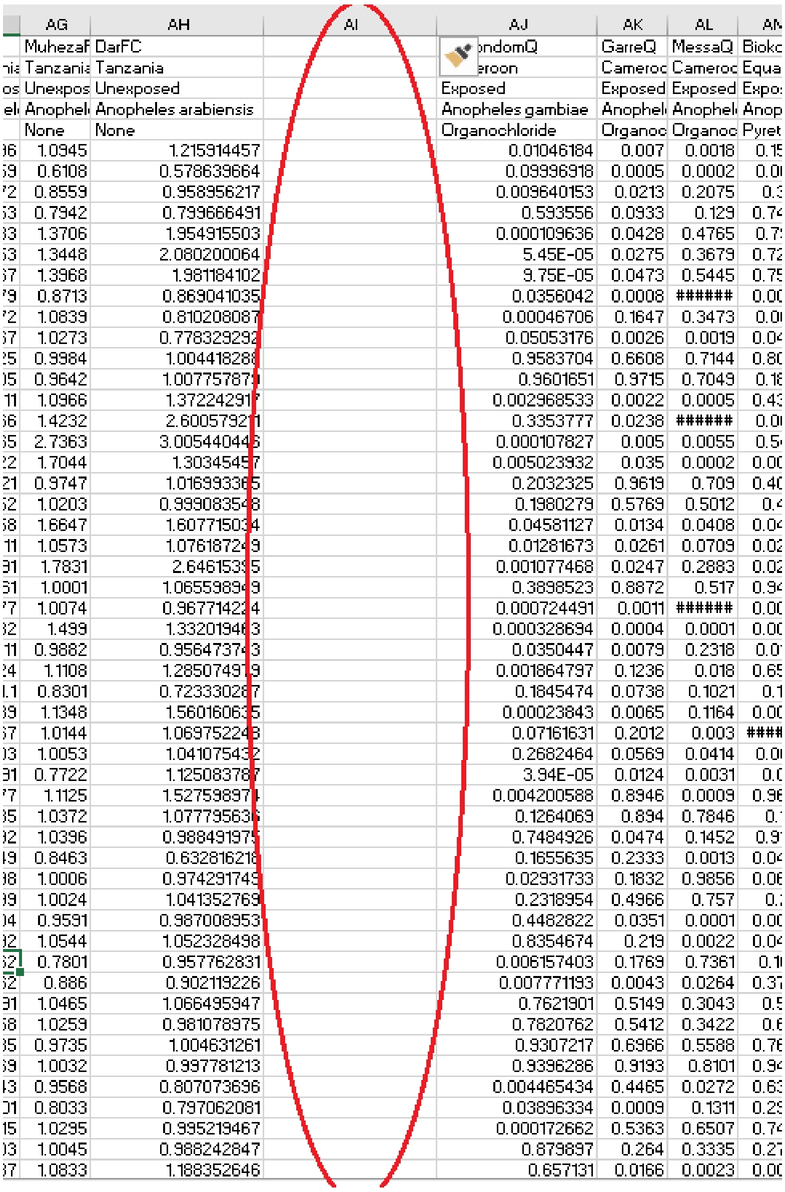
Adding data to the web-based app

To add new datasets to the existing web-based app, please email the first author with the new experimental files and designs, in addition to latitude and longitude of collection site of the resistant population: victoria.ingham@lstmed.ac.uk.

Adding data to a local IR-TEx installation

The following is a step-by-step guide to adding data to a **local installation** of IR-TEx. Please follow the steps below

1. Get the data files - within the github repository, there are a number of files, including the user guide. Please download the following files and save then to an appropriate folder: Fold Changes.txt, geography.txt, IR-TEx.R.
2. Open the file Fold Changes.txt – this is a large excel file with RAW fold changes (NOT logged).
3. Examine the file - The first row contains the name of the population, followed by FC (necessary) and also by Q (necessary). Find the final column with FC and insert a row to the right, as illustrated below:



	AG	AH	AI	AJ	AK	AL	AM
	MuhezaF	DarFC		ondomQ	GarreQ	MessaQ	Bioko
nia	Tanzania	Tanzania		eroon	Cameroo	Cameroo	Equa
os	Unexpos	Unexposed		Exposed	Exposed	Exposed	Expo:
el	Anophel	Anopheles arabiensis		Anopheles gambiae	Anophel	Anophel	Anop
	None	None		Organochloride	Organoc	Organoc	Pyret
36	1.0945	1.215914457		0.01046184	0.007	0.0018	0.15
59	0.6108	0.578639664		0.09996918	0.0005	0.0002	0.0
72	0.8559	0.958956217		0.009640153	0.0213	0.2075	0.3
53	0.7942	0.799666491		0.593556	0.0933	0.129	0.74
33	1.3706	1.954915503		0.000109636	0.0428	0.4765	0.7
53	1.3448	2.080200064		5.45E-05	0.0275	0.3679	0.72
57	1.3968	1.981184102		9.75E-05	0.0473	0.5445	0.75
79	0.8713	0.869041035		0.0356042	0.0008	######	0.00
72	1.0839	0.810208087		0.00046706	0.1647	0.3473	0.0
57	1.0273	0.778329292		0.05053176	0.0026	0.0019	0.04
25	0.9984	1.004418288		0.9583704	0.6608	0.7144	0.80
05	0.9642	1.007757879		0.9601651	0.9715	0.7049	0.18
11	1.0966	1.372242917		0.002968533	0.0022	0.0005	0.43
36	1.4232	2.600579211		0.3353777	0.0238	######	0.0
35	2.7363	3.005440446		0.000107827	0.005	0.0055	0.5
22	1.7044	1.303454567		0.005023932	0.035	0.0002	0.00
21	0.9747	1.016993365		0.2032325	0.9619	0.709	0.40
52	1.0203	0.999083548		0.1980279	0.5769	0.5012	0.4
58	1.6647	1.607715034		0.04581127	0.0134	0.0408	0.04
11	1.0573	1.076187249		0.01281673	0.0261	0.0709	0.02
31	1.7831	2.646153395		0.001077468	0.0247	0.2883	0.02
61	1.0001	1.065598949		0.3898523	0.8872	0.517	0.94
77	1.0074	0.967714224		0.000724491	0.0011	######	0.00
32	1.499	1.332019463		0.000328694	0.0004	0.0001	0.00
11	0.9882	0.956473743		0.0350447	0.0079	0.2318	0.0
24	1.1108	1.285074919		0.001864797	0.1236	0.018	0.65
1.1	0.8301	0.723330287		0.1845474	0.0738	0.1021	0.1
39	1.1348	1.560160635		0.00023843	0.0065	0.1164	0.00
57	1.0144	1.069752248		0.07161631	0.2012	0.003	####
03	1.0053	1.041075432		0.2682464	0.0569	0.0414	0.0
31	0.7722	1.125083787		3.94E-05	0.0124	0.0031	0.0
77	1.1125	1.527598974		0.004200588	0.8946	0.0009	0.96
35	1.0372	1.077795636		0.1264069	0.894	0.7846	0.
92	1.0396	0.988491975		0.7484926	0.0474	0.1452	0.9
49	0.8463	0.632816218		0.1655635	0.2333	0.0013	0.04
98	1.0006	0.974291749		0.02931733	0.1832	0.9856	0.06
39	1.0024	1.041352769		0.2318954	0.4966	0.757	0.
04	0.9591	0.987008953		0.4482822	0.0351	0.0001	0.00
92	1.0544	1.052328498		0.8354674	0.219	0.0022	0.04
52	0.7801	0.957762831		0.006157403	0.1769	0.7361	0.1
52	0.886	0.902119226		0.007771193	0.0043	0.0264	0.37
31	1.0465	1.066495947		0.7621901	0.5149	0.3043	0.5
58	1.0259	0.981078975		0.7820762	0.5412	0.3422	0.6
35	0.9735	1.004631261		0.9307217	0.6966	0.5588	0.76
59	1.0032	0.997781213		0.9396286	0.9193	0.8101	0.94
43	0.9568	0.807073696		0.004465434	0.4465	0.0272	0.63
01	0.8033	0.797062081		0.03896334	0.0009	0.1311	0.29
15	1.0295	0.995219467		0.000172662	0.5363	0.6507	0.74
03	1.0045	0.988242847		0.879897	0.264	0.3335	0.27
37	1.0833	1.188352646		0.657131	0.0166	0.0023	0.00

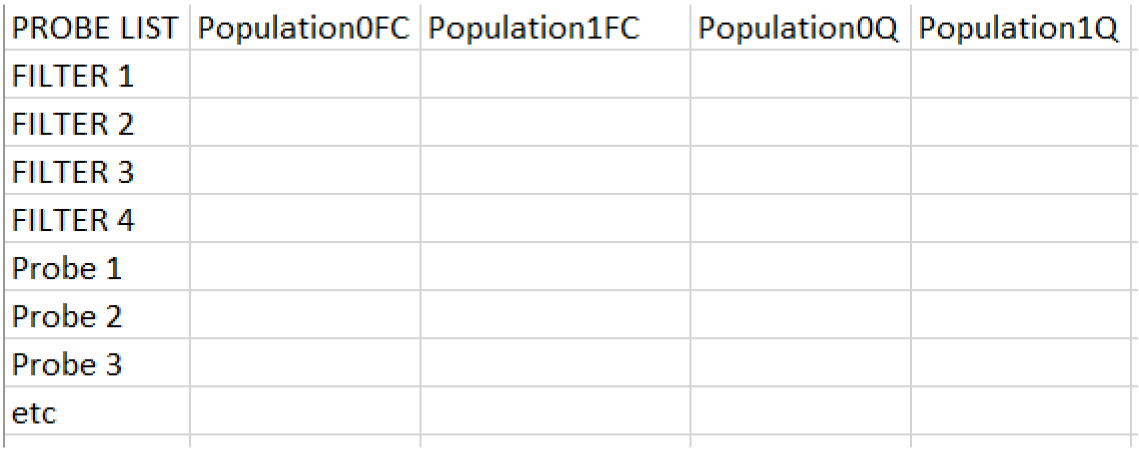
4. Insert dataset descriptors - In the top cell insert a name for the population followed by FC, in the second the country of the resistant population, in the third either Exposed or Unexposed dependent upon whether the resistant population is exposed, in the fourth the species Anopheles gambiae, Anopheles coluzzi or Anopheles arabiensis (MUST be full species ID) and in the fifth cell down the class of insecticide or ‘None’ if unexposed. Underneath this paste the raw fold changes corresponding to the probe of this row. Repeat this with Q values in the furthest right column on the sheet, keeping all information in the top 5 columns identical whilst replacing FC with Q.
5. Mapping data - Now open geography.txt, it will contain a column of resistant population names, exactly how they appear in the Q value columns of Fold Changes.txt, a latitude and a longitude. Input your population name EXACTLY as it appears in the Q value column under the last row of the first column, followed by the latitude and longitude of the collection site (or approximate original location) of the new resistant population in the dataset.
6. Install – replace the existing files with your newly modified files (Fold Changes.txt, geography.txt) and restart the application.

Section 3: Adapting the App to Handle Other Expression Data

Part of the utility of this app is use in a wider field than insecticide resistance alone. It will specifically be useful in fields that have a variety of transcriptomic data from different experiments, from which there will be some merit to analyse them together. To achieve this, there will need to be changes to the key code. For the purposes of this walkthrough, no previous knowledge of R is necessary but to fully adapt the code, R knowledge will be required. Due to this, the walkthrough will cover inputting data with ONLY 4 filtering criteria.

Creating a new data file

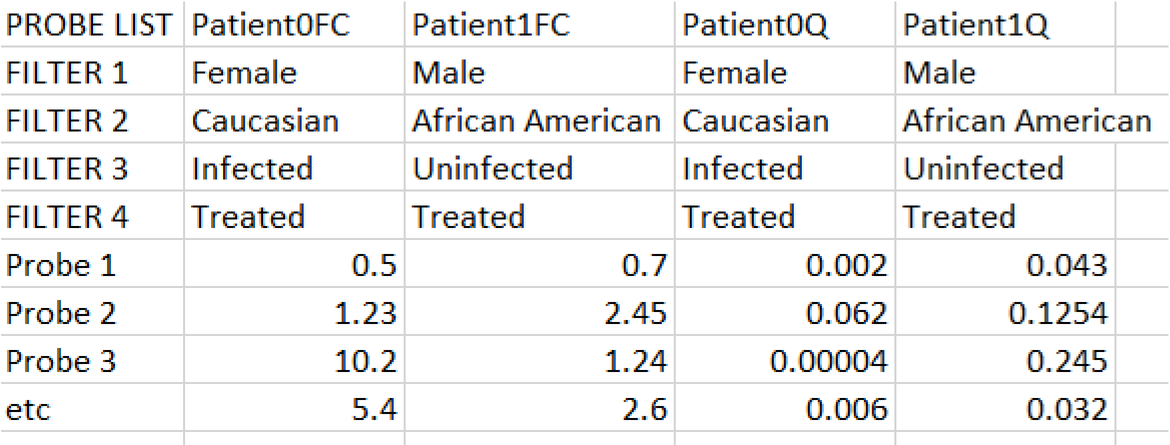
The first task is to create a new Fold Changes.txt file (tab delimited) to appropriately match the template provided for insecticide resistance as seen below.



PROBE LIST	Population0FC	Population1FC	Population0Q	Population1Q
FILTER 1				
FILTER 2				
FILTER 3				
FILTER 4				
Probe 1				
Probe 2				
Probe 3				
etc				

Populating the data file

Enter the appropriate parameters for each filter, as in the example below. Capitalisation and punctuation is important in R so make sure everything is consistent and DO NOT change, for example, between ‘female’ and ‘Female’. These filters should overlap otherwise you will not be able to select multiple datasets



PROBE LIST	Patient0FC	Patient1FC	Patient0Q	Patient1Q
FILTER 1	Female	Male	Female	Male
FILTER 2	Caucasian	African American	Caucasian	African American
FILTER 3	Infected	Uninfected	Infected	Uninfected
FILTER 4	Treated	Treated	Treated	Treated
Probe 1	0.5	0.7	0.002	0.043
Probe 2	1.23	2.45	0.062	0.1254
Probe 3	10.2	1.24	0.00004	0.245
etc	5.4	2.6	0.006	0.032

Modifying the R code to accept your data

Once the data is in with the probes matching across rows for all datasets, the R code can now be modified by following the steps below.

The guide will assume that geography is NOT relevant to these datasets, so the map will be removed, as well as geography.txt. If geography is relevant ignore step a.) and the final step deleting lines 479-584 and modify the geographical parameters as in section (ii).

a. Open the R code in a text editor, delete the line *library(dismo)* and the line *geography<-read.delim('geography.txt',header=T)*
b. *titlePanel('IR-TEx', windowTitle = 'IR-TEx')* Change the name within the ' to your own dataset ie *titlePanel('Patient Data', windowTitle = 'Patient Data')*
c. *textInput('textInput','Transcript ID',value='AGAP008212-RA')* Change the AGAP008212-RA to a probe on your array, ie in the above example *textInput('textInput','Transcript ID',value='Probe 1')*. It must be exact.
d. *checkboxGroupInput('CountryInput','Select Relevant Countries',c('Burkina Faso','Cote D`Ivoire','Cameroon','Equatorial Guinea','Zambia','Tanzania','Sudan','Uganda','Togo'),selected=c('Burkina Faso','Cote D`Ivoire','Cameroon','Equatorial Guinea','Zambia','Tanzania','Sudan','Uganda','Togo'))* Change this line to your appropriate filters and those you wish to be selected when the app opens. For the example above this would become *checkboxGroupInput('CountryInput','Select Relevant Sex',c('Male','Female'),selected=c('Male','Female'))* This must correspond to the first row of filters, don't change 'CountryInput'
e. checkboxGroupInput('ExposureInput','Select Exposure Status',c('Exposed','Unexposed'),selected=c('Exposed','Unexposed')) Change this to correspond to filter 2 checkboxGroupInput('ExposureInput','Select Ethnicity Status',c('Caucasian','African American'),selected=c('Caucasian','African American')) This must correspond to the second row of filters, don't change 'ExposureInput'
f. checkboxGroupInput('SpeciesInput','Select Relevant Species',c('Anopheles gambiae','Anopheles coluzzi','Anopheles arabiensis'),selected = c('Anopheles coluzzi')) Change this to correspond to filter 3 checkboxGroupInput('SpeciesInput','Select Infectious Status',c('Infected','Uninfected'),selected = c('Infected','Uninfected')) This must correspond to the third row of filters, don't change 'SpeciesInput'
g. checkboxGroupInput('InsecticideInput','Select Insecticide Class',c('Pyrethroid','Organochloride','Carbamate','None'),selected = c('Pyrethroid','None')) Change this to correspond to filter 4 checkboxGroupInput('InsecticideInput','Select Treatment Regime',c('Treated','Untreated'),selected = c('Treated','Untreated')))) This must correspond to the fourth row of filters, don't change 'InsecticideInput'

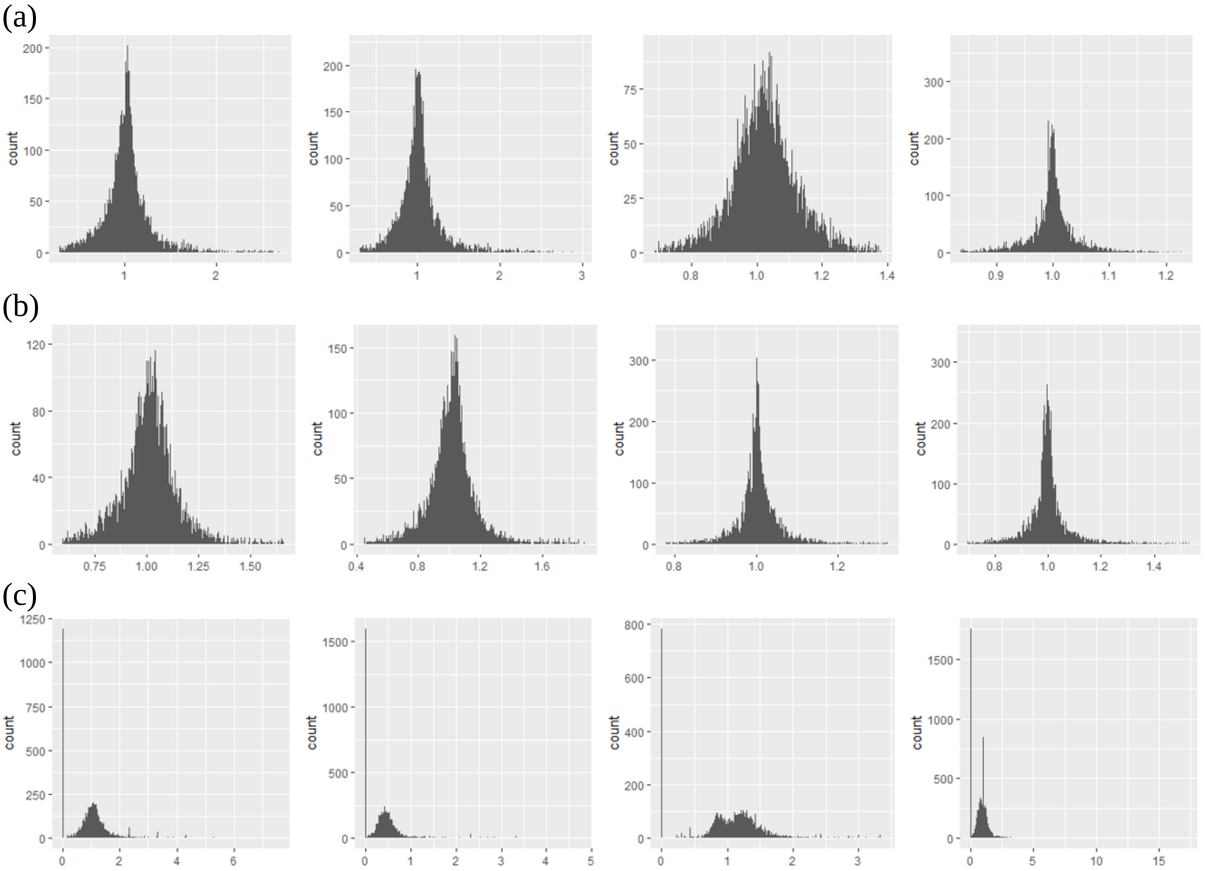
There can be as many filters as required, each must be surrounded by 'X' and separated by a comma. The above correspond to lines 1-47 of the code. We will now scroll to lines 479-584. These will be deleted and begin/end to:

```
START: output$Geography <- renderPlot({
END    paste("Significant Transcripts Only (p", as.expression("<="),"0.05): FC > 5 = Red, FC > 1 = Amber, FC < 1 = Green",sep="")
  })
```

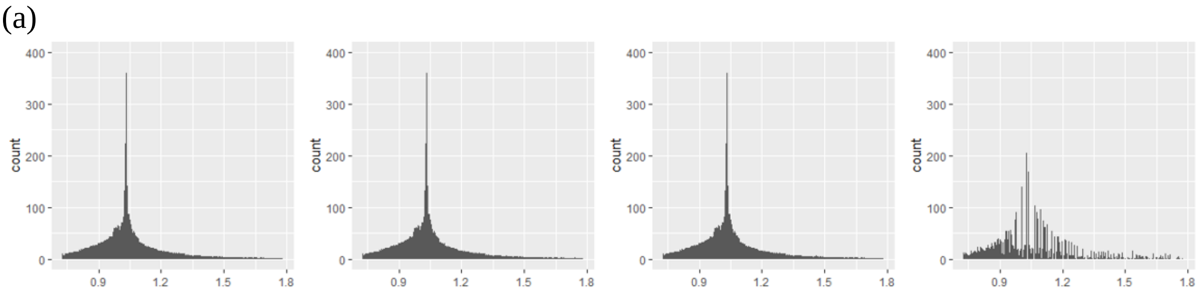
Inclusive of the brackets: })

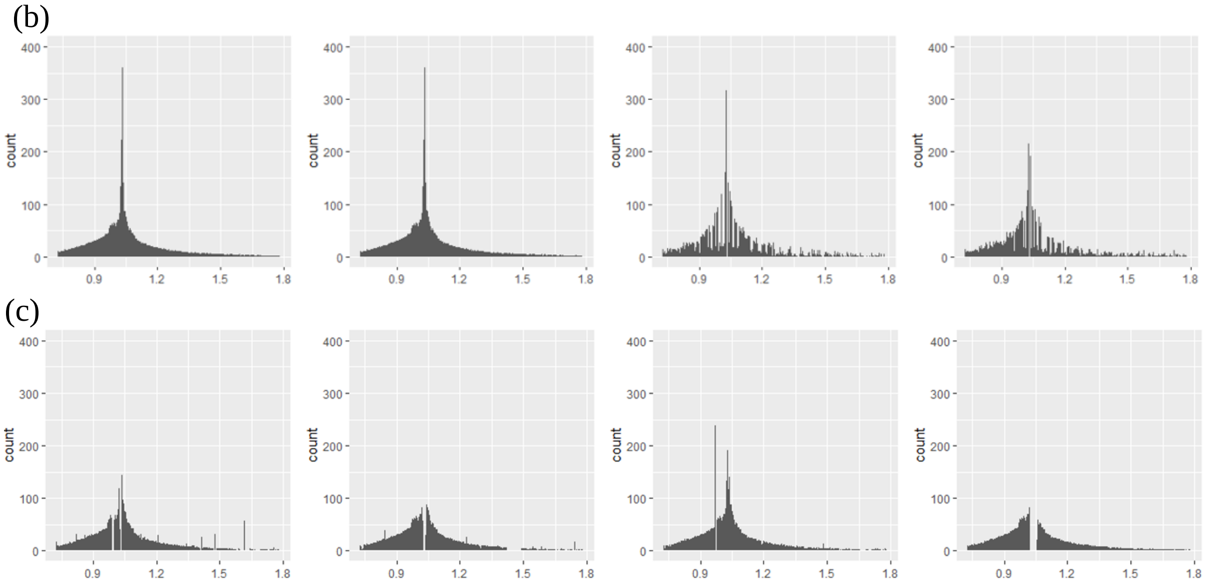
Use of one colour arrays and RNASeq data with the app

One-colour arrays have often been used due to the original high price of two-colour arrays. Similarly, different arrays and different analysis techniques are often used on array data. With the advent of RNAseq data, this leads to a further confounding variable when using this app. Each of these techniques have different distributions of fold change, below are examples of the fold change distributions of four different experiments from (a) two colour arrays, (b) one colour arrays and (c) RNAseq. In each case fold change is on the x axis and the y axis represents the number of transcripts for each fold change.



These differences in distributions are a confounding factor when integrating this data, specifically with use of correlation networks which are driven by extreme values, as often seen in RNAseq. To input data, follow guide in section ii, once this is complete and the data stored in a tab delimited text file, normalisation can be performed. By using a quantile normalisation across a table containing all fold changes, it is possible to make the distributions identical in terms of statistical properties and is a technique widely employed with microarray data analysis. As seen below, these distributions are much more similar after normalisation; this transformation can also be used across array platforms (Affymetrix, Agilent, Ilumina etc.)





These data are kindly contributed by Duo Peng of the Catteruccia lab, Harvard T.H Chan School of Public Health.

Section 4: References

For further information about algorithms and a description of the methods and use case please see our publication "**Transcriptomic meta-signatures identified in *Anopheles gambiae* populations reveal previously undetected insecticide resistance mechanisms**" V.A Ingham, S. Wagstaff and H. Ranson. Nature Comms.

Where to find help

To install IR-TEx, you will need to download the current file from https://github.com/LSTMScientificComputing/IR-TEx and execute it in a ShinyR environment.

Detailed instructions how to deploy the ShinyR environment can be found on the Shiny project webpage - https://shiny.rstudio.com. Example instructions on how to configure ShinyR for Ubuntu can be found here.

Example tutorial on installing for Ubuntu 16.04
https://www.digitalocean.com/community/tutorials/how-to-set-up-shiny-server-on-ubuntu-16-04

Example tutorial on installing for Ubuntu 14.04
https://www.digitalocean.com/community/tutorials/how-to-set-up-shiny-server-on-ubuntu-14-04

Version

SW, VAI, HR authored the user guide
Data for different -omics platforms provided by DP
Created on: 20th August 2018
Updated on: 19th October 2018
Version: 1.1