VectorBase Hands-on Workshop
August 16, 2015
Intercontinental Hotel
Cali, Colombia

VectorBase Insecticide Resistance (IR) data submission

In this workshop exercise, we will learn about VectorBase's current data submission format for population biology data, with an emphasis on submitting insecticide resistance data.

VectorBase uses a widely adopted spreadsheet format for 'omics data annotation and submission called ISA-Tab. You can read about it here. Please look carefully at the overview diagram.
http://isatab.sourceforge.net/format.html (or Google: isa-tab format)

For VectorBase PopBio submissions, each ISA-Tab file should only contain one study. Thus the ISA-Tab concepts of Investigation and Study correspond to the PopBio Project and its list of Samples, while ISA-Tab Assays correspond directly to PopBio Assays.

In this exercise we are going to look at an example, partially completed ISA-Tab spreadsheet and add some fictional data to it.

First open the spreadsheet, log in to google if you haven't already and use the File menu to "Make a copy…"
https://goo.gl/JXNUCG

After you have made a copy, you will be able to edit the spreadsheet.

Ignore any **greyed out text** in the spreadsheets. These parts are ignored by VectorBase.

Yellow backgrounds indicate parts that should be filled in as you read through this worksheet.

Make sure you have seen the tabs at the bottom of your screen - we will be switching between the different sheets a lot!

The investigation sheet/tab

This contains information about the project/publication/dataset. Note that ontology terms are used throughout ISA-Tab spreadsheets, including the "Study design descriptors" section.

Tasks:

1. Add the correct EFO accession number for "Compound treatment design" using the OntoMaton plugin, following the instructions in the spreadsheet.
2. Add information about a publication in the relevant section. If you wanted to add two publications you would add the second in column C of the sheet.

The "Study assays" section outlines which broad types of assay have been performed (the four same assay types that you saw in the search tutorial) and which spreadsheet tabs contain information about them. Note that "a_collection.txt" refers to the spreadsheet tab "a_collection".

The tab prefixes are as follows

- i = investigation
- s = study/samples
- a = assays
- p = phenotypes
- g = genotypes (not used in this example)

Note also the **four protocols** that are described here. They will be referenced from the three assay sheets. Note that they are described by ontology terms and optional free text. The IR assay has three parameters defined here. We'll see those in action later.

Samples

The samples sheet should only contain **incontrovertible facts** about samples, such as sex and developmental stage. Species information has its own assay sheet (see below).

The "Source Name" column is not used in VectorBase. You must put something here but it doesn't matter what.

Comment columns appear in the PopBio Browser with the subheading you provide in the square brackets. Comments can also be used to provide link-outs to external databases (talk to a VectorBase curator for details).

Task: add a my-sample-2 row to this sheet, copying all the information from the row above, as appropriate. If you like, you can change "female" to "male" and look up the correct PATO ontology accession number using the OntoMaton lookup tool.

Ontology term doubts?

When in doubt, type the text for the term you want and ask a VectorBase curator to help you find the correct ontology or add a new term to an ontology as appropriate.

You can share your Google spreadsheet with the curator and he/she can help you fill it in.

In fact, a curator will probably set up the spreadsheet for you to complete by adding the data rows only.

In all ISA-Tab sample and assay sheets, you can add any "Characteristics [xxx]" columns you like, as long as the "xxx" is a valid ontology term.

Field collection

Since this is the first assay sheet, we'll point out some basics.

The "Sample Name" in the first column should be the same as one of the samples in the sample sheet. The "Assay Name" should be derived from the sample name such that each distinct assay has a distinct "Assay Name" (see the species and IR_WHO sheets for examples). The "Protocol REF" should refer to the "Study Protocol Name" in the investigation sheet. We use capital letters here so that they stand out, but this is not mandatory.

Task: add a second row to describe the collection of your "my-sample-2". To keep things simple, use the same collection protocol as the first row. For the column "Characteristics [Collection site (VBcv:0000831)]" (column M), you should consult the Gazetteer (GAZ) ontology in the VectorBase ontology browser. GAZ isn't currently available via OntoMaton. If you can't find the exact town or village you are looking for, see if you can find a wider region or district (or failing that, use the country GAZ term) - then enter more detailed information (without ontology terms) in the peach-coloured columns. Leave the peach-coloured columns blank if you found the exact GAZ term you were looking for. For latitude/longitude coordinates, you can right click on Google Maps ("what's here?") to get some real-world coordinates.

Species identification assays

Switch to the **a_species** sheet/tab.

In VectorBase PopBio we like to know *how* the species of a sample was determined. Therefore, we don't just ask you to fill in a "species column" in a spreadsheet. Instead, you should provide information about each species identification assay performed, even if that was just a visual identification.

VBsp can be browsed and searched via "Tools->Ontology browser", and then selecting "VectorBase CV" (CV stands for controlled vocabulary).

This sheet shows how a single sample can have *two different assays* performed on it. Note that the "Assay Name" column is generated by a formula to make sure each assay has a unique name.

Task: enter species terms and accession numbers as directed in the spreadsheet.

Insecticide resistance assays

Change to the "**a_IR_WHO**" tab.

> **ISA-Tab extensions**
> Assay sheets (e.g. "a_IR_WHO") are formally ISA-Tab format.
>
> Phenotype sheets (e.g. "p_IR_WHO") are not officially ISA-Tab but they are ISA-Tab-like. Some columns, e.g. Observable and Attribute, are not allowed in official ISA-Tab.

This is another assay sheet, and again note how the Assay Name column is generated from information within the sheet that can change between assays. (If assays with different concentrations of the same insecticide were performed, then these would also need to have different Assay Names).

Tasks:

1. Add a row for my-sample-1 tested for **deltamethrin** resistance (you can choose any concentration, and perhaps choose a different unit?).
2. Add rows for my-sample-2, for assays with both insecticides.

You should end up with four data rows in this sheet.

Insecticide resistance phenotypes

Change to the "**p_IR_WHO**" tab.

This is where the *results* of the assays are recorded.

Tasks:

1. Extend the formula in cell A2 down three more rows. It should then show the four Assay Names from the previous assay sheet.
2. Do the same for the "Phenotype Name" column
3. Copy rows C-N down too, but change the Value column to some suitable fictional values.

Watch out for unwanted auto-increment on numbers when you copy-down spreadsheet rows (you don't want this to happen to ontology accession numbers).

Wrap up

You have now provided (fictional) data in the correct format for VectorBase PopBio submission for two samples assayed with two insecticides each, including the field collection and species identification metadata for these samples. We hope that you would find it easy to add further samples and assays.

We would very much like to hear your questions and comments on this process. Please type and send your message via VectorBase contact form: https://www.vectorbase.org/contact