Vector Mosquito Monitoring Via Biodiversity Specimen/DNA Data Sharing

Best Practices Guide

Global Mosquito Alert Consortium

2021-04-07

Contents

1 Introduction

The Global Mosquito Alert Consortium's (GMAC's) best practices guides offer experiences gained from a variety of projects that use citizen science to better understand and combat disease-vector mosquitoes. The goal is to create a growing repository of information about how best to use and customize the GMAC's citizen science toolkit for local implementation.

The present guide encompasses Pillar 4 of the GMAC toolkit: Vector Mosquito Monitoring via biodiversity specimen/DNA data sharing. This is a shared document to be used for developing an effective international citizen science network for vector mosquito monitoring, this part covering best practice in collecting mosquito specimens and DNA. Best practice for research data sharing has been handled differently by different fields, but an attempt to synthesis these was carried out by the Toronto International Data Release Workshop in 2009 (see: http://www.nature.com/nature/journal/v461/n7261/full/461168a.html). DNA data sharing has a particularly long history, and has lead the way in much of biological research. The principles for rapid release of genome-sequence data were first formulated the Bermuda (1996-1998) and Fort Lauderdale (2003) meetings, and remain the accepted policies recognised by the genomics community, journals and funders. Data producers were obliged to submit their data to one of the three INSDC databases (EBI, NCBI or DDBJ) as soon as possible after the assembled sequence has met a set of quality evaluation criteria. In exchange for 'early release' of their data, the international sequencing centers retained the right to be the first to describe and analyze their complete datasets in peer-reviewed publications. Leading to the Toronto workshop, attendees endorsed the value of rapid pre-publication data release for large reference datasets in biology and medicine that have broad utility and agreed that pre-publication data release should go beyond genomics. This asks data producers to release their data, but they request a protected time period to allow them to be the first to publish the data set, this should be limited to global analyses of the data and ideally expire within one year.

Beyond making data open, best practice is now to follow the FAIR Guiding Principles for scientific data management and stewardship (see: https://www.nature.com/articles/sdata201618) which focuses on the interoperability and machine readability of datasets. To maximise the re-use of this citizen collected data the Global Mosquito Alert Consortium would ideally follow these overarching principles and practices. For more practical purposes there are more field specific guidelines and systems for dealing with collecting this type of insect biodiversity data that will be outlined in more detail below.

2 Pillar Objectives

- Collected specimens will be submitted to their local operational partner via postal mail. Users should receive notifications when their specimens arrive, and when new vectors are identified in their area.
- DNA is extracted and then sequenced using standard protocols and practices to enable identification of the species.
- Pictures (facultative) of the specimens collected may also be stored alongside other metadata such as location, sampling details and other notes (such as number of species in an area).
- If the specimens are not destroyed in the process of extracting the DNA, and if the DNA is not all used up in the process of sequencing it then these would be ideally stored for future use, and sample/voucher number details stored and made public.
- Sequencing data and the contextual information around it should be made publicly available following community norms and practices to maximise its research value.

3 Citizen Scientist Roles and Motivations

How are citizen scientists involved?

- They find specimens, send them to the local operational partner, and contribute in tracking these public health hazards.

How are tasks divided between citizen scientists and experts?

- See above.

What motivates the citizen scientists?

- Fighting disease. Personal health and safety for their families and communities.
- They can be the first to discover invasive or even completely new species.
- Part of classroom education projects to boost understanding of biodiversity, public health, the consequences of climate change, and increase genetic literacy.

4 Ethics

Unless human DNA or pictures are accidentally passed on there should be no ethical considerations with this type of data.

5 Data Collection

Metadata should be collected in a manner that is GBIF compliant and ideally follows the Darwin Core Standard (DwC) that offers a stable, straightforward and flexible framework for compiling biodiversity data from varied and variable sources. https://www.gbif.org/darwin-core. Would be best practice to include these fields in the data collection process (e.g. people sending specimens asked to include this in a webform, app or printed checklist).

If this is too complicated, the eBOL Community Web Portal provides guidelines and resources for educators, students and citizen scientists seeking to organize, manage, and share their project data with the broader DNA barcoding community in a form that is compatible with research standards. This also includes DNA Barcoding Assistant app for smartphones that helps streamline and standardize the collection of sample information by students http://www.educationandbarcoding.org/DNAB.php

Additional Citizen friendly guidelines for DNA barcoding are available here: http://www.dnabarcoding101.org/resources/

See pillar 1 for best practice on collecting the associated images.

6 Data Processing and Validation

How is data processed once it is collected?

There are a number of different sequencing technologies and approaches (e.g. barcoding, genotyping, whole genome sequencing) that can be used to collect this data, and they are also changing rapidly so it is probably beyond the scope of this document to go into detail here. The output of all of these technologies are quite standardised DNA sequence formats. They then ideally need to be used in GBIF compliant ways.

How is it validated?

Darwin core validation tool https://tools.gbif.org/#dwca-validator

7 Data Presentation and Use

How is data presented to the public and to specific end users (e.g. public health agencies)?

One good example of how data can be presented to the public is Wildlife-of-our-Homes. This project allowed contributing citizens who collected microbe samples from their homes to “see” and interact with the data using a platform built from open source visualisation tools http://robdunnlab.com/projects/wild-life-of-our-homes/data-visualization/

The BOLD systems-SDP (student data portal) is an integrated workbench supporting the assembly, analysis and publication of DNA barcode data by students http://v3.boldsystems.org/index.php/SDP_Home For researchers and Public Health officials, GBIF provides easy and fast global access to data and metadata, and maps to browse geographically.

What are the particular needs of these end-users?

Following the FAIR principles, for biodiversity data best practice would be to have the data interoperable and indexable by GBIF, so their policies regarding formatting, metadata and open licensing will need to be followed. This means CC0 or CC-BY licenses need to be followed.

The other policy to be careful to follow is the “Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity”.

How is data used for vector management and research?

Submitting the data to GBIF means it can be integrated and easily viewed on the global biodiversity occurrence map for all researchers and public health workers to see and track with quite high confidence.

8 Data Structures and Repository Links

If possible, what are the core requirements for data structure?

See above and the tools section for details on GBIF and sequencing community norms.

What repositories should data from this pillar be sent to?

Raw sequencing data is universally mandated go in one of the three INSDC databases (NCBI, EBI, DDBJ), but there is more scope in what is done with the downstream processed data if you want to provide bulk access or use another database.

Barcoding data should ideally be deposited with Barcode of Life (http://www.boldsystems.org/), but this is brokered via the INSDC databases anyway.

9 Existing Tools

What concrete tools can be drawn from to implement this pillar?

- GBIF use Darwin core metadata and there are templates for the collection of this https://tools.gbif.org/#templates
- Best practice for sequencing metadata is encapsulated by MIxS checklists: http://www.ebi.ac.uk/ena/submit/mixs-checklists
  - Of which barcoding data is covered: http://www.ebi.ac.uk/ena/submit/species-barcode-checklist
- GBIF collects geographically tagged INSDC sequences, which as processed by the INSDC databases (inc EBI) will follow this, e.g.:
- https://www.gbif.org/dataset/ad43e954-dd79-4986-ae34-9ccdbd8bf568
- If new species are detected they need to be registered (http://zoobank.org/) in Zoobank and the procedure needs to follow ICZN guidelines (http://www.iczn.org/)
- For students and citizen scientists the eBOL Community Web Portal includes barcoding tools and apps to collect and analyse their samples in standardised, research friendly ways http://www.educationandbarcoding.org/

10 Case Studies

Mueckenatlas

The specific data associated with the collected mosquito(es), such as collection date, locality, description of the collection site, weather etc., is supposed to be fed into the German national mosquito database CULBASE where collection data from various German monitoring programmes and research projects are gathered. https://mueckenatlas.de/

All the focus seems to be keeping this for the research community, so this is not really an example of data sharing in citizen science.

Invasive Mosquito Project

The data (mosquitoes) are gathered by teachers and citizens and sent to local mosquito control/public health agencies or in their absence sent to the United States Department of Agriculture. The information is publically shared and the mosquito samples are stored at the USDA for future population genetics projects. http://www.citizenscience.us/imp/.

Kissing bug citizen science program

Researchers in Texas created a kissing bug citizen science program to educate the public about Chagas disease, and to create a mechanism for the public to submit the triatomine ‘kissing bug’ vectors.

http://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0004235

Nice example of educating and interacting with the public. Didn't collect sequencing data though.

Serendipitous discovery of Wolbachia genomes in multiple Drosophila species

Making raw sequencing data of drosophila species available without restriction in NCBI allowed other researchers several years later to discover completely new wolbachia symbiont genomes in them.

https://doi.org/10.1186/gb-2005-6-3-r23

Example of both the importance of collecting raw sequencing data and of data-driven research. Open data can enable serendipitous future discoveries that the original data producers may not have even thought about.

Open sourcing genomes / crowdsourcing killer outbreaks

Releasing the first sequencing data from the deadly 2011 German E. coli outbreak CC0 and in a citeable manner with data DOIs to give the data producers credit helped kick-start a burst of crowd-sourced, curiosity-driven analyses from bioinformaticians (and at least one citizen blogger) around the world. Doing science in this accelerated way sped up diagnosis and treatments, within a couple of a days a potential ancestral strain had been identified, and enabled the rapid development of diagnostic tests and anti-microbial agents.

http://opendatahandbook.org/value-stories/en/open-sourcing-genomes/

This example has been used as an example for EU science policy, with the Royal Society in the UK using it as an example of "the power of intelligently open data", and highlighting it on the cover of their influential "Science as an Open Enterprise" report.