Variation search for VectorBase species

Table of Contents

Introduction

This tutorial concentrates on Advanced Search for molecular level variation data such as Single Nucleotide Polymorphisms (SNPs), small insertion/deletions events (indels) and large scale structural rearrangement such as the 2La inversion event seen in *Anopheles gambiae*. These variation data have been derived from publications, analyses supplied to VectorBase by external teams, and projects such as the *Anopheles* 1000 genomes program (Ag1000G).

Variation data are stored in a number of different VectorBase resources depending on the nature of the data and how they are displayed. Population level studies and phenotypic data are stored within the PopBio system, while molecular level data relating to defined DNA and protein variations and their linked phenotypes are stored within species specific variation databases. These databases are linked to the genome browsers and Sample and Genotype Explorer tools allowing you to view the variation data in its genomic context, the studies from the data was derived, and compare variations at loci between different species. The VectorBase advanced search facility extends these capabilities, allowing you to search for variation data across all of these resources and to link back to the relevant display tools.

There are over 100 million SNP variants available across all of the VectorBase species, the vast majority of which lie in repetitive or uncharacterized regions of the genome. To shrink the search space to a manageable level VectorBase indexes variants that can be directly linked to annotated genes. This reduces the number of SNP variants to just under 11 million entries across 16 species (as of August 2018). While the variation search does not index all possible variants, we believe that the majority of our users will be interested in variants which are linked to genes and have predictable consequences to coding sequences. It is still possible to access all 100 million variants through the genome browsers and REST API programmatic interface, and since the variation search facility provides direct links from variants back to the genome browsers it always possible to navigate to regions of genomic interest and view all of the available variation data. We also provide links to VCF files for the original study data which contain all of the reported variants. If you wish to work with large amounts of variant data we recommend that you download the relevant VCF files for your purposes.

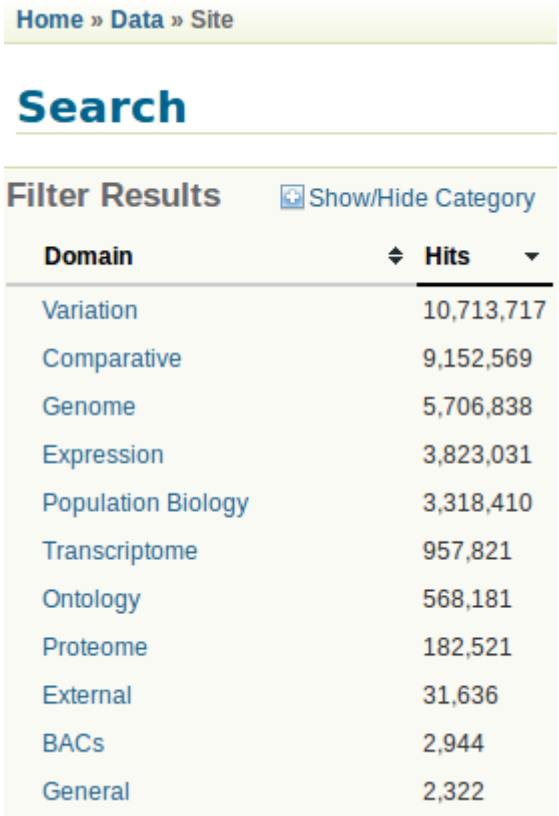
Locating variation data via site search

Variation data can be accessed by selecting the “Advanced search” option below the site search box at the top right corner of all VectorBase site pages.

Fig 1. Location of VectorBase search tool.

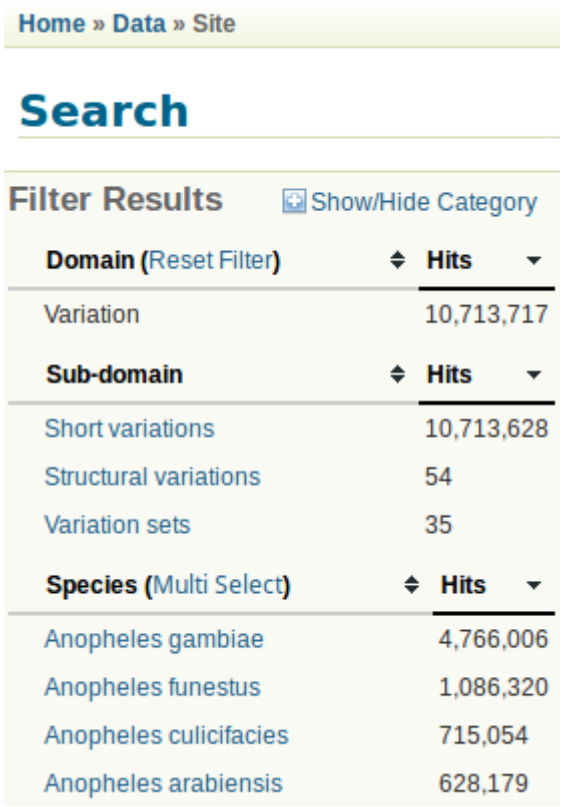
Selecting the “Advanced search” facility will show a list of ‘domains’ for different types of data available at VectorBase.

Fig 2. Data domains available via VectorBase search.



Selecting the “Variation” domain allows you to view more information about the structure of the sub-domain showing counts of short variations, structural variations and variation sets (collections of variants from a study).

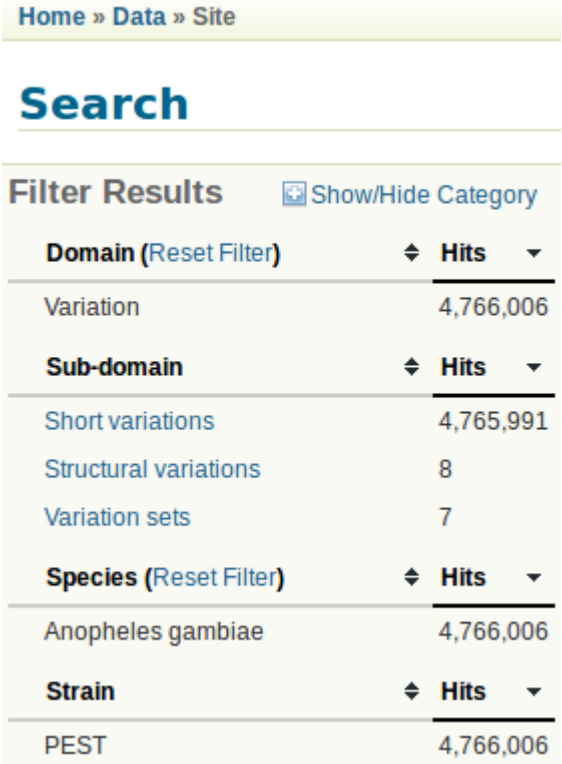
Fig 3. Variation domain and subdomain.



Below the sub-domain the total numbers of variant counts for each species are also shown. This is a useful way to quickly gain an idea of how much variation data is available for the species you are interested in.

To view the variation data for a species in detail simply select it from the list.

Fig 4. Restricting variation view by species.



To view the contents of a sub-domain, select the relevant link. The three sub-domains currently available are:

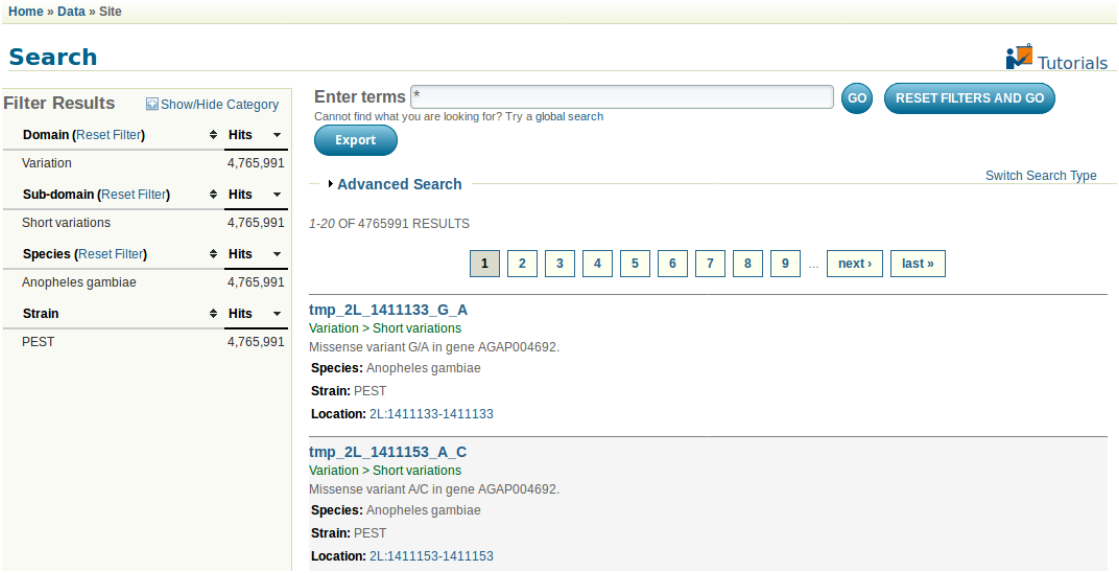
1. Short variations - SNPs and small insertion/deletion events (< 1 Kbp in size).
2. Structural variations - large scale genomic rearrangements in the range of 1Kbp -> 10 Mbp
3. Variation sets - collections of variant data from a study

Examples and descriptions of each of the subdomains are shown on the following pages.

Short variants

Short variants include SNPs and small insertion/deletion events (< 1 Kbp in size). Some example records are shown below,

Fig 5. Example short variant data record displayed in search.



Each record shows a brief description of the variant including information such as the species and strain the variant was observed in, which gene it is associated with, the possible effect on the coding sequence, and the genomic location. Clicking on the location or variant name hyperlinks will take you to the corresponding location or record in the genome database.

Fig 6. Variant record displayed in the genome browser.

Advanced search

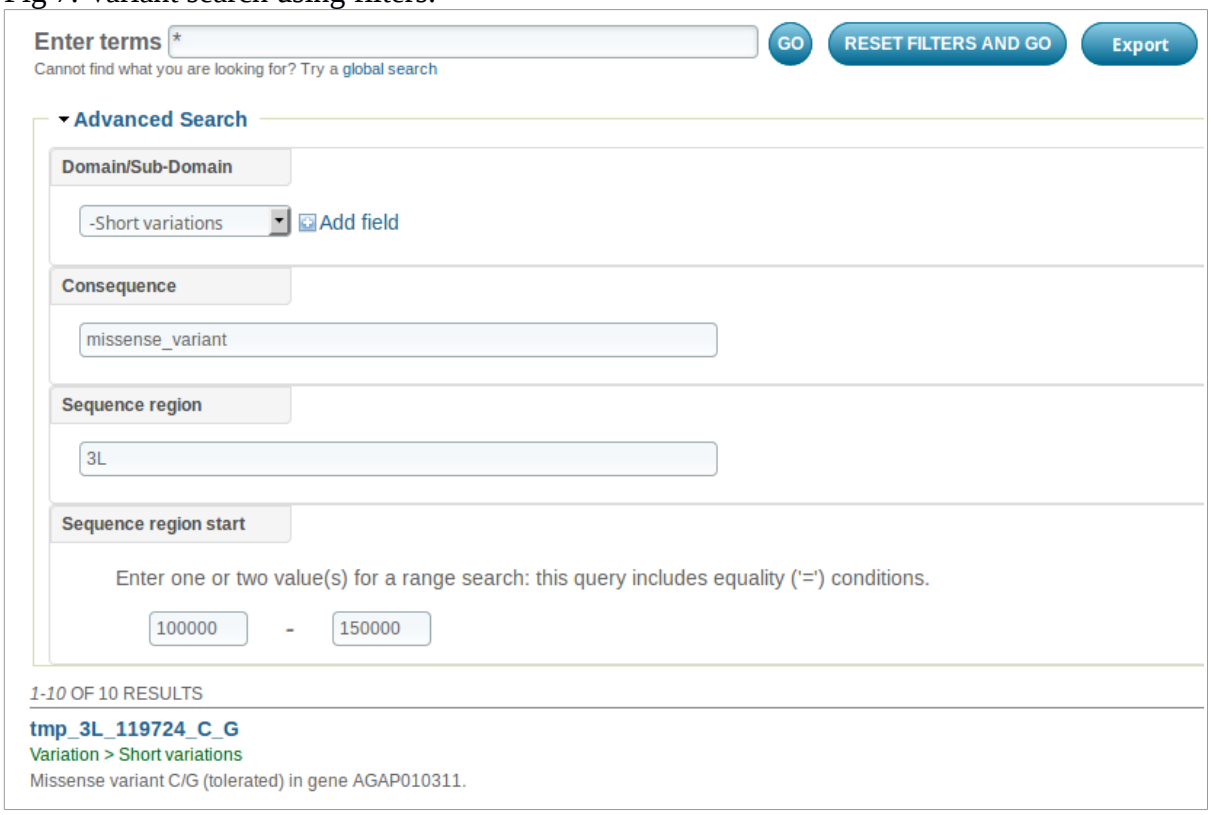
Selecting the "Advanced Search" option displays a list of fields to filter the current list of entries.

The available fields allow to search in different ways:

- Using a genomic location range to locate variants: if you want to retrieve a single SNP, only use the "Sequence region" and the "Sequence region start" fields. (NB: a SNP has identical start and stop coordinates.)
- Consequences of the variant (consequence_types for genomic consequences, and SIFT for proteins).
- Variants associated with a given phenotype.
- Variants found in a given gene or transcript.
- Variants from a given Popbio project.

All fields can use autocomplete.

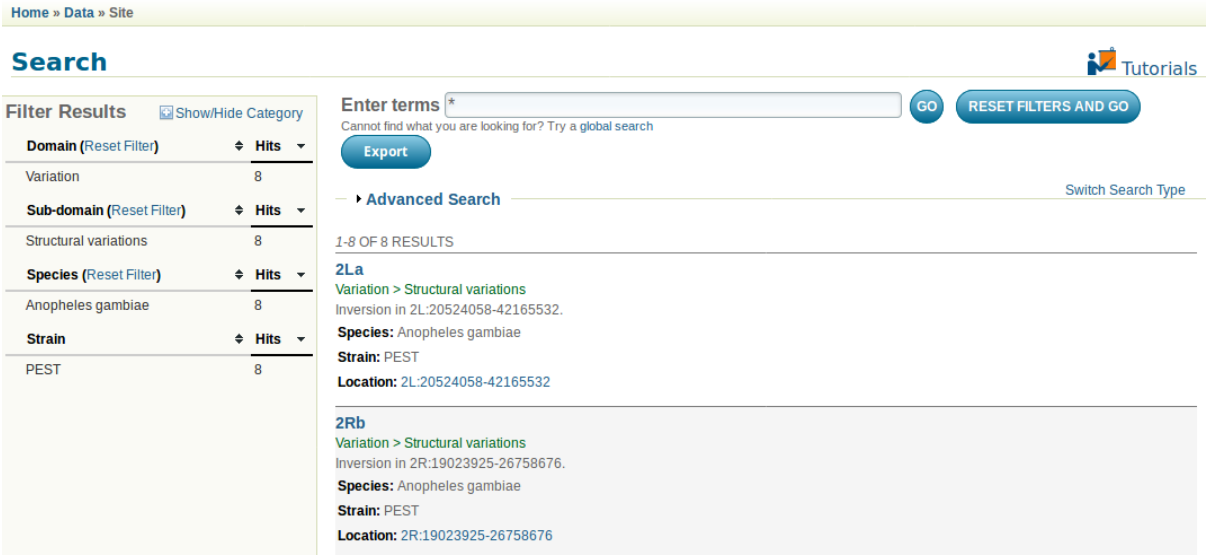
Fig 7. Variant search using filters.



Structural variations

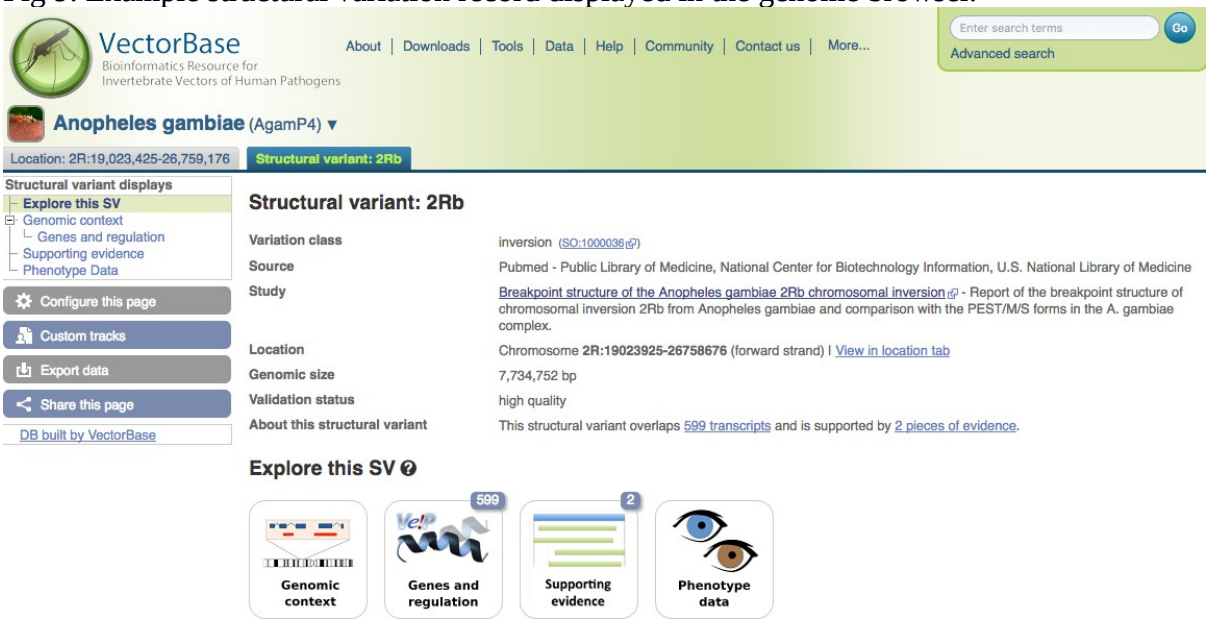
Structural variations represent large scale genomic rearrangements in the range of 1Kbp -> 10 Mbp. Examples of structural variations include the *Anopheles gambiae* 2La inversion, or copy number variations (CNV) for genes involved in insecticide resistance.

Fig 8. Example structural variation record displayed in search.



Each of these structural variation records contains an active link back to the genome browser which allows you to view summary data about the genomic location, genes within this interval, supporting evidence and phenotypic data

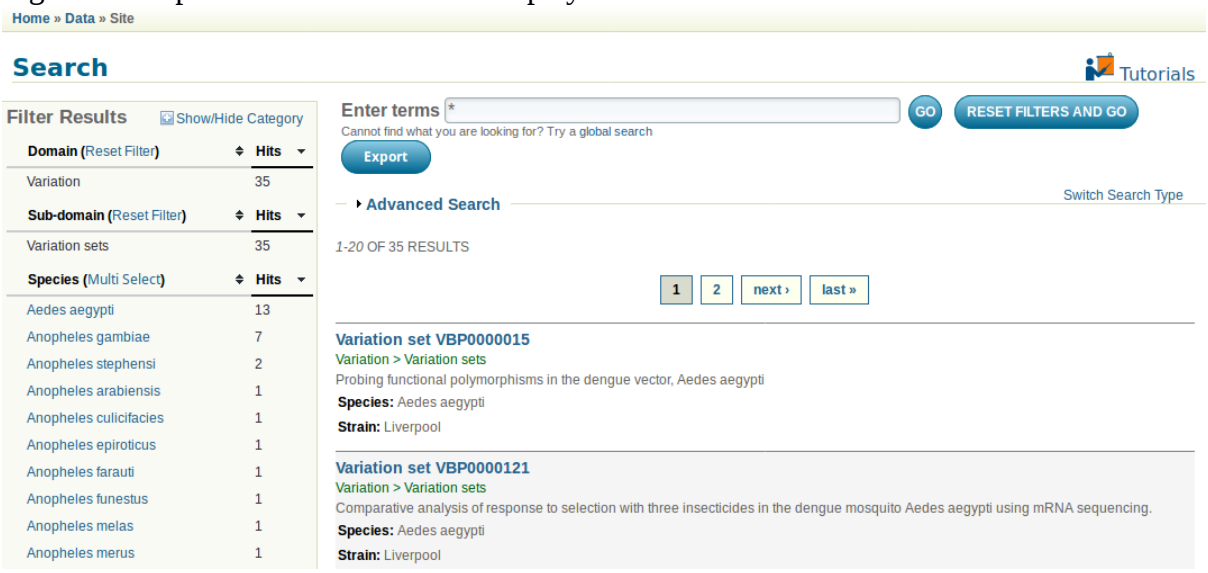
Fig 9. Example structural variation record displayed in the genome browser.



Variation sets

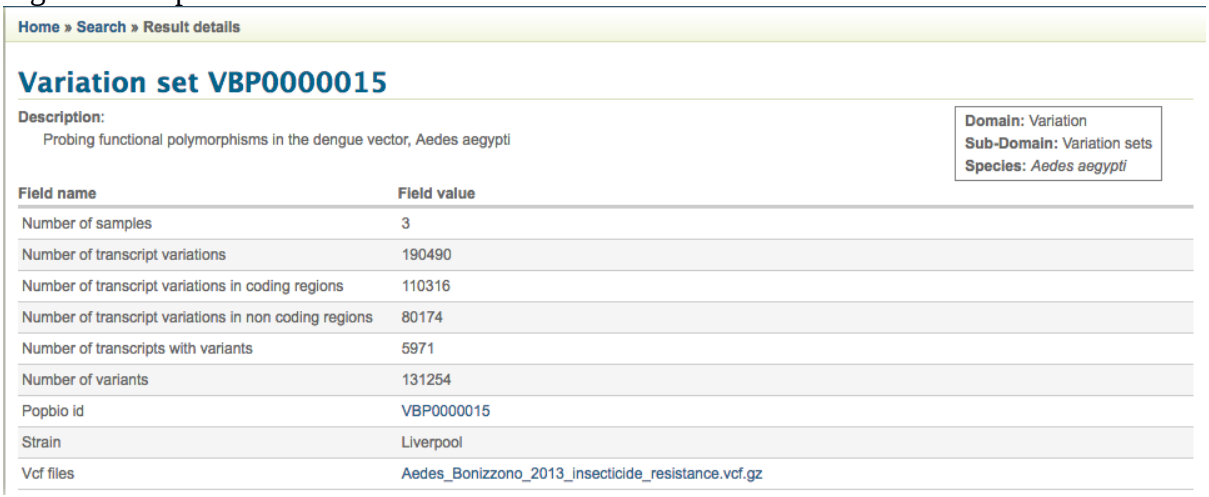
Variation sets are collections of variant data from a study. Each study has a corresponding PopBio database record indicated by the VBP record id (see example below).

Fig 10. Example variation set records displayed in search.



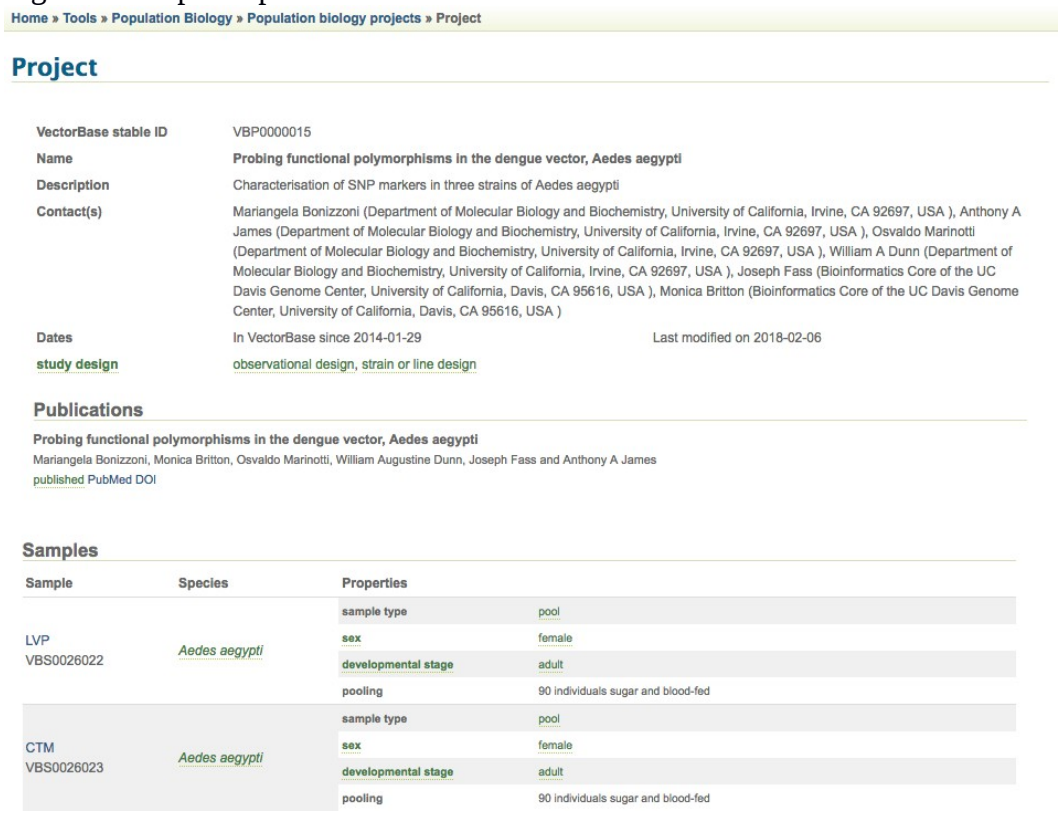
If you follow the link for each variation set you will be presented with a summary describing the number of samples in the study, statistics about the number of variants present, and a link to the full PopBio database record (shown below).

Fig 11. Example variation set record in detail.



Within this record are links to the full PopBio record for the study (Fig 11) and the VCF file(s) for this study containing the genotype calls from this study. Most studies will have only a single VCF file, but some larger studies may have multiple files. VCF files are typically large and are compressed as gzip files to speed up downloading. You will need to use a file decompression tool such as unzip or gunzip to work with these files after download.

Fig 12. Example PopBio record.



Home » Tools » Population Biology » Population biology projects » Project

Project

VectorBase stable ID	VBP0000015
Name	**Probing functional polymorphisms in the dengue vector, Aedes aegypti**
Description	Characterisation of SNP markers in three strains of Aedes aegypti
Contact(s)	Mariangela Bonizzoni (Department of Molecular Biology and Biochemistry, University of California, Irvine, CA 92697, USA), Anthony A James (Department of Molecular Biology and Biochemistry, University of California, Irvine, CA 92697, USA), Osvaldo Marinotti (Department of Molecular Biology and Biochemistry, University of California, Irvine, CA 92697, USA), William A Dunn (Department of Molecular Biology and Biochemistry, University of California, Irvine, CA 92697, USA), Joseph Fass (Bioinformatics Core of the UC Davis Genome Center, University of California, Davis, CA 95616, USA), Monica Britton (Bioinformatics Core of the UC Davis Genome Center, University of California, Davis, CA 95616, USA)
Dates	In VectorBase since 2014-01-29 — Last modified on 2018-02-06
study design	observational design, strain or line design

Publications

Probing functional polymorphisms in the dengue vector, Aedes aegypti
Mariangela Bonizzoni, Monica Britton, Osvaldo Marinotti, William Augustine Dunn, Joseph Fass and Anthony A James
published PubMed DOI

Samples

Sample	Species	Properties	
LVP VBS0026022	*Aedes aegypti*	**sample type**	pool
		sex	female
		developmental stage	adult
		pooling	90 individuals sugar and blood-fed
CTM VBS0026023	*Aedes aegypti*	**sample type**	pool
		sex	female
		developmental stage	adult
		pooling	90 individuals sugar and blood-fed

Downloading variation data sets

To download the full variation set for a study click on the VCF files link(s). This will trigger a download of a compressed (gzip) Variation Call Format (VCF) file. VCF files can be very large (>10GB) so it may take some time to download a complete data set. VCF files for VectorBase species may have come from a variety of sources, and will contain varying amounts of metadata regarding samples, call analysis, quality scores etc.

The VCF format is an evolving standard and you can find more documentation at the following sites:

http://www.internationalgenome.org/wiki/Analysis/variant-call-format
https://github.com/samtools/hts-specs
https://en.wikipedia.org/wiki/Variant_Call_Format

Feedback and help

We are always interested in receiving feedback, suggestion for further development, or providing help. Please contact us either via the website contact page

https://www.vectorbase.org/contact

or email directly to

info@vectorbase.org