

# Variation search for VectorBase species

## Table of Contents

Introduction.....	1
Locating variation data via site search.....	2
Short variants.....	5
Structural variations.....	7
Variation sets.....	8
Downloading variation data sets.....	10
Feedback and help.....	10

## [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).

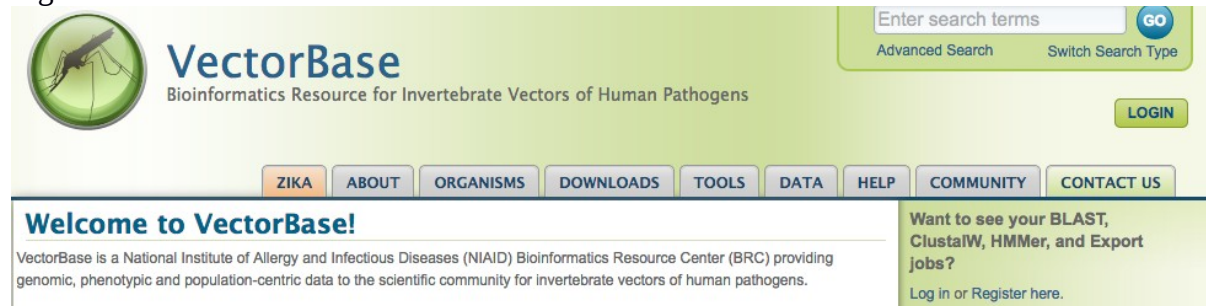
Variation data are stored in a number of different VectorBase resources depending on the nature of the data and how they are displayed. Population level studies and phenotypic data are stored within the PopBio system, while molecular level data relating to defined DNA and protein variations and their 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.

[Home](#) » [Data](#) » [Site](#)

## Search

**Filter Results** [Show/Hide Category](#)

Domain	Hits
Variation	10,713,717
Comparative	9,152,569
Genome	5,706,838
Expression	3,823,031
Population Biology	3,318,410
Transcriptome	957,821
Ontology	568,181
Proteome	182,521
External	31,636
BACs	2,944
General	2,322

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.

[Home](#) » [Data](#) » [Site](#)

## Search

**Filter Results**  Show/Hide Category

<b>Domain (Reset Filter)</b>	<b>Hits</b>
Variation	10,713,717
<b>Sub-domain</b>	<b>Hits</b>
Short variations	10,713,628
Structural variations	54
Variation sets	35
<b>Species (Multi Select)</b>	<b>Hits</b>
Anopheles gambiae	4,766,006
Anopheles funestus	1,086,320
Anopheles culicifacies	715,054
Anopheles arabiensis	628,179

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.

[Home](#) » [Data](#) » [Site](#)

## Search

---

**Filter Results** [Show/Hide Category](#)

<b>Domain (Reset Filter)</b>	<b>Hits</b>
Variation	4,766,006
<b>Sub-domain</b>	<b>Hits</b>
<a href="#">Short variations</a>	4,765,991
<a href="#">Structural variations</a>	8
<a href="#">Variation sets</a>	7
<b>Species (Reset Filter)</b>	<b>Hits</b>
<a href="#">Anopheles gambiae</a>	4,766,006
<b>Strain</b>	<b>Hits</b>
<a href="#">PEST</a>	4,766,006

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.

Home » Data » Site

### Search

Filter Results  Show/Hide Category

Domain (Reset Filter)	Hits
Variation	4,765,991
Sub-domain (Reset Filter)	Hits
Short variations	4,765,991
Species (Reset Filter)	Hits
Anopheles gambiae	4,765,991
Strain	Hits
PEST	4,765,991

Enter terms

Cannot find what you are looking for? Try a global search

Advanced Search

1-20 OF 4765991 RESULTS

...

**tmp\_2L\_1411133\_G\_A**  
Variation > Short variations  
Missense variant G/A in gene AGAP004692.  
**Species:** Anopheles gambiae  
**Strain:** PEST  
**Location:** 2L:1411133-1411133

**tmp\_2L\_1411153\_A\_C**  
Variation > Short variations  
Missense variant A/C in gene AGAP004692.  
**Species:** Anopheles gambiae  
**Strain:** PEST  
**Location:** 2L:1411153-1411153

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.

Variant displays

- Explore this variant
- Genomic context
- Genes and regulation
- Flanking sequence
- Genotype frequency
- Phenotype data
- Sample genotypes
- Linkage disequilibrium
- Phylogenetic context
- Citations

DB built by VectorBase

### tmp\_2L\_1411133\_G\_A SNP

Most severe consequence **missense variant** | See all predicted consequences **G/A**

Alleles **G/A**

Location **Chromosome 2L:1411133 (forward strand) | VCF: 2L 1411133 tmp\_2L\_1411133\_G\_A G A**

HGVS names **This variant has 6 HGVS names - Hide**

- 2L:g.1411133G>A
- AGAP004692-RA.1:c.-223+17G>A
- AGAP004692-RB.1:c.13G>A
- AGAP004692-PB.1:p.Ala5Thr
- AGAP004692-RC.1:c.-215+92G>A
- AGAP004692-RD.1:c.-54+17G>A

Original source **Sequencing studies of MR4 derived insect colonies**

About this variant **This variant overlaps 4 transcripts and has 1 sample genotype.**

### Explore this variant

- Genomic context
- Genes and regulation
- Flanking sequence
- Population genetics
- Phenotype data
- Sample genotypes
- Linkage disequilibrium
- Phylogenetic context
- Citations
- View variant in Ag1000G browser

### Using the website

- Video: [Browsing SNPs and CNVs in Ensembl](#)
- Video: [Clip: Genome Variation](#)
- Video: [BioMart: Variation IDs to HGNC Symbols](#)
- Exercise: [Genomes and SNPs in Malaria](#)

### Reference materials

- [Variation Quick Reference card](#)

### Analysing your data

Test your own variants with the Variant Effect Predictor

## [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.

The screenshot shows a web interface for advanced variant search. At the top, there is a search bar with the text "Enter terms \*" and a "GO" button. Below the search bar, there are two buttons: "RESET FILTERS AND GO" and "Export". A message below the search bar reads "Cannot find what you are looking for? Try a [global search](#)".

The main section is titled "Advanced Search" and contains four filter categories:

- Domain/Sub-Domain:** A dropdown menu is set to "-Short variations" with an "Add field" button next to it.
- Consequence:** A text input field contains "missense\_variant".
- Sequence region:** A text input field contains "3L".
- Sequence region start:** A text input field contains "100000", followed by a hyphen "-", and another text input field containing "150000". Below this, a note reads: "Enter one or two value(s) for a range search: this query includes equality ('=') conditions."

At the bottom of the interface, it displays "1-10 OF 10 RESULTS" and a result entry: **tmp\_3L\_119724\_C\_G**, with a link "Variation > Short variations" and the text "Missense variant C/G (tolerated) in gene AGAP010311."

## 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.

The screenshot shows a search interface with a 'Filter Results' sidebar on the left and search results on the right. The sidebar includes filters for Domain (Variation: 8), Sub-domain (Structural variations: 8), Species (Anopheles gambiae: 8), and Strain (PEST: 8). The search results section shows two entries: 2La and 2Rb. Each entry includes a link to 'Structural variations', the genomic location (e.g., 2L:20524058-42165532), and the species and strain information.

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.

The screenshot shows the 'Structural variant: 2Rb' record in the genome browser. The page includes a navigation menu at the top, a search bar, and a sidebar with options like 'Explore this SV', 'Genomic context', 'Genes and regulation', 'Supporting evidence', and 'Phenotype Data'. The main content area displays details for the 2Rb inversion, including its location (2R:19023925-26758676), genomic size (7,734,752 bp), and validation status (high quality). It also provides a link to the 'Breakpoint structure of the Anopheles gambiae 2Rb chromosomal inversion' and mentions that the variant overlaps 599 transcripts and is supported by 2 pieces of evidence. At the bottom, there are four interactive icons for 'Genomic context', 'Genes and regulation', 'Supporting evidence', and 'Phenotype data'.

## 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.

The screenshot shows a search interface with a left sidebar for filtering results and a main content area for search results. The sidebar includes filters for Domain, Sub-domain, and Species. The main content area shows search results for 'Variation set VBP0000015' and 'Variation set VBP0000121'. The search bar contains the text 'Enter terms' and 'Export' button. The results are displayed in a list format with pagination controls.

Home » Data » Site

### Search

Enter terms

Cannot find what you are looking for? Try a global search

Switch Search Type

Filter Results  Show/Hide Category

Domain (Reset Filter)	Hits
Variation	35

Sub-domain (Reset Filter)	Hits
Variation sets	35

Species (Multi Select)	Hits
Aedes aegypti	13
Anopheles gambiae	7
Anopheles stephensi	2
Anopheles arabiensis	1
Anopheles culicifacies	1
Anopheles epiroticus	1
Anopheles farauti	1
Anopheles funestus	1
Anopheles melas	1
Anopheles merus	1

1-20 OF 35 RESULTS

1 2 next › last »

**Variation set VBP0000015**  
Variation > Variation sets  
Probing functional polymorphisms in the dengue vector, Aedes aegypti  
**Species:** Aedes aegypti  
**Strain:** Liverpool

**Variation set VBP0000121**  
Variation > Variation sets  
Comparative analysis of response to selection with three insecticides in the dengue mosquito Aedes aegypti using mRNA sequencing.  
**Species:** Aedes aegypti  
**Strain:** Liverpool

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.

The screenshot shows a detailed view of a variation set record. It includes a description, a table of field values, and a summary box with domain, sub-domain, and species information. The table lists various statistics such as the number of samples, transcript variations, and variants.

Home » Search » Result details

## Variation set VBP0000015

Description:  
Probing functional polymorphisms in the dengue vector, Aedes aegypti

Domain:	Variation
Sub-Domain:	Variation sets
Species:	Aedes aegypti

Field name	Field value
Number of samples	3
Number of transcript variations	190490
Number of transcript variations in coding regions	110316
Number of transcript variations in non coding regions	80174
Number of transcripts with variants	5971
Number of variants	131254
Popbio id	VBP0000015
Strain	Liverpool
Vcf files	Aedes_Bonizzono_2013_insecticide_resistance.vcf.gz



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

---

<b>VectorBase stable ID</b>	VBP0000015
<b>Name</b>	<b>Probing functional polymorphisms in the dengue vector, <i>Aedes aegypti</i></b>
<b>Description</b>	Characterisation of SNP markers in three strains of <i>Aedes aegypti</i>
<b>Contact(s)</b>	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 )
<b>Dates</b>	In VectorBase since 2014-01-29 <span style="float: right;">Last modified on 2018-02-06</span>
<b><a href="#">study design</a></b>	<a href="#">observational design, strain or line design</a>

---

## 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	<i>Aedes aegypti</i>	sample type: pool
		sex: female
		developmental stage: adult
		pooling: 90 individuals sugar and blood-fed
CTM VBS0026023	<i>Aedes aegypti</i>	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](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](mailto:info@vectorbase.org)