

VectorBase Variation

(includes six demos with videos)

Gareth Maslen & Gloria I. Giraldo-Calderón
August 2018

Adapted from 'Variation data in Ensembl and the Ensembl VEP'
by Erin Haskell

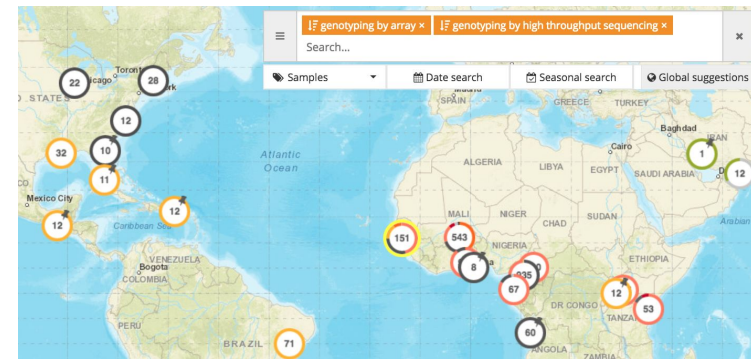


VectorBase

Bioinformatics Resource for Invertebrate Vectors of Human Pathogens

Objectives

1. Types of VectorBase variation data
2. Present sample queries to find variants in:
 - genes and regions
 - different geographic locations
3. Analyse your own variation data for potential effects on genes
 - Variant Effect Predictor, VEP



Types of VectorBase variation data

- SNPs
- INDELS
- Microsatellites
- Structural variants:
A. gambiae 2La inversions,
A. aegypti CNVs

Data display:

- GB: genome browser
- PopBio: population biology map
- Search (coming soon)

Data source:

Scientific publications
and databases

MalariaGEN
GENOMIC EPIDEMIOLOGY NETWORK

Ag1000G

Variation Data

Summary of available variation data by organism

Reference species	Assembly	SNP calls	Indel calls
<i>Aedes aegypti</i> Liverpool	AaegL3	332,445	4,081
<i>Aedes aegypti</i> LVP_AGWG	AaegL5	313,612	1,761
<i>Anopheles arabiensis</i> Dongola	AaraD1	10,164,339	1,022,752
<i>Anopheles culicifacies</i> A-37	AculA1	9,154,354	900,063
<i>Anopheles epiroticus</i> Epiroticus2	AepiE1	3,281,528	267,653
<i>Anopheles farauti</i> FAR1	AfarF2	7,226,209	871,433
<i>Anopheles funestus</i> FUMOZ	AfunF1	12,920,105	987,886
<i>Anopheles gambiae</i> PEST	AgamP4	59,236,057	2,168,425
<i>Anopheles melas</i> CM1001059_A	AmelC2	3,677,409	418,730
<i>Anopheles merus</i> MAF	AmerM2	6,053,485	558,680
<i>Anopheles minimus</i> MINIMUS1	AminM1	4,206,083	222,877
<i>Anopheles quadriannulatus</i> SANGWE	AquaS1	10,137,655	931,982
<i>Anopheles sinensis</i> SINENSIS	AsinS2	5,841,551	409,698
<i>Anopheles stephensi</i> SDA-500	AsteS1	5,856,684	574,778
<i>Anopheles stephensi</i> Indian	Astel2	366,367	0
<i>Biomphalaria glabrata</i> BB02	BglaB1	10,031,395	0
<i>Culex quinquefasciatus</i> Johannesburg	CpipJ2	2	0
<i>Ixodes scapularis</i> Wikel	IscaW1	1,776,352	0

SNPs and INDELS:

- 18 genomes
- 16 species (there are two *A. aegypti* and two *A. stephensi*)

Microsatellites:

- *Aedes aegypti* samples worldwide

Reference and Alternative alleles

Reference genome/alleles: GACTAAATGCATCG

frequency G= 0.05, frequency T=0.95

T is the allele in all Diptera (flies and mosquitoes)

G causes insecticide resistance

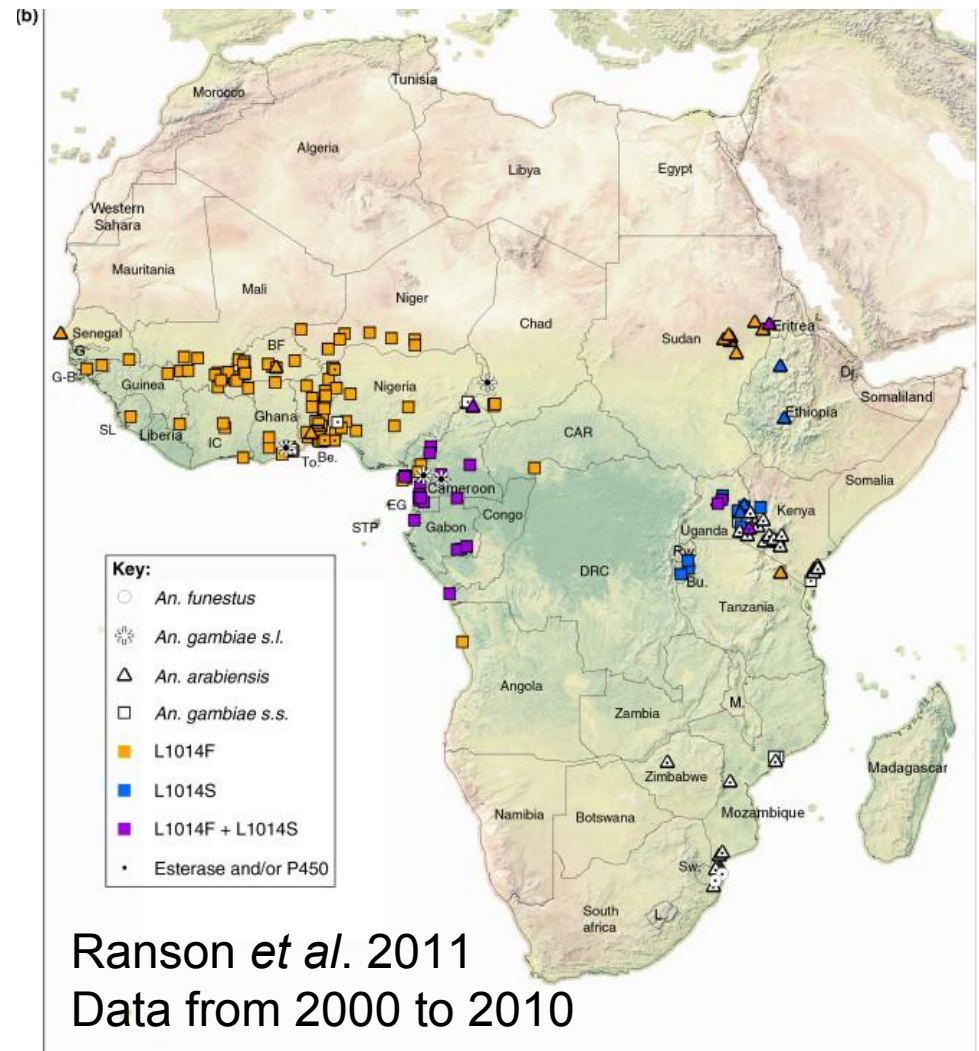
G is the allele in the contig used in the reference genome

- G is the reference allele
- T is the alternative allele
- Alleles are reported as G/T (reference/alternative)

The reference allele is not necessarily the wild type!

Reference and Alternative alleles: **Sample case**

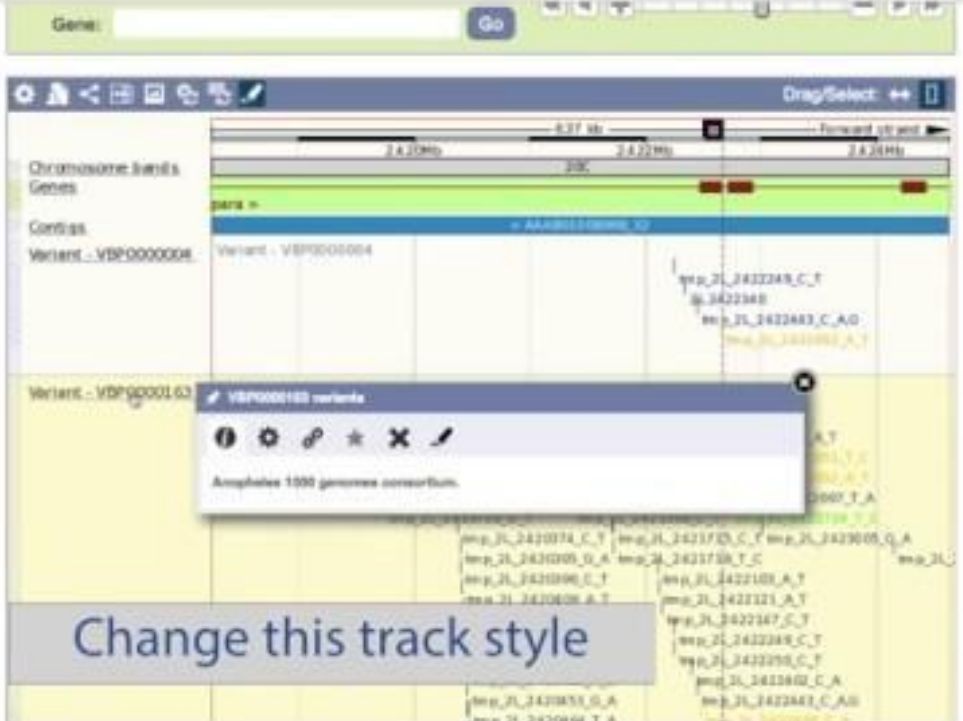
- *Musca domestica* aa residue 1014 (origin of nomenclature)
- L1014: susceptible (wild type)
- L1014F & L1014S: Alleles conferring resistance to pyrethroids
- *A. gambiae* aa residue 980 (2L:2422652, gene AGAP004707)



Query #1: Are there SNPs for the gene AGAP004707?

The image displays a genomic alignment viewer with a DNA sequence on the left and variant information on the right. The sequence is: `RACAAATTTACTGAACTAACAAATTCAAATAGACATTTGAAATRAACTACTTGGCCGTGA`. The right side lists variants such as `35591:tmp_Z1_2362118_A_C`. A semi-transparent text box in the lower-left corner contains the text: "They are all the variants for this gene available in VectorBase".

Query #2: How to visualize the SNPs in the gene and its splice variants?



The image shows a screenshot of the Ensembl genome browser interface. At the top, there is a search bar labeled "Gene:" with a "Go" button. Below this, a navigation bar includes icons for home, back, forward, and search, along with a "Drag/Select" button. The main area displays a genomic track for a gene, with a scale from 24,270kb to 24,280kb. A green bar represents the gene structure, and a blue bar below it shows the gene ID "AA020119.1". A variant track is visible, showing a variant "Variant - VEP0000004". A tooltip for this variant is open, displaying a list of SNPs with their coordinates and alleles, such as "rs_21_2422248_C_T" and "rs_21_2422248_C_A". A text box at the bottom of the screenshot says "Change this track style".

Track styles:

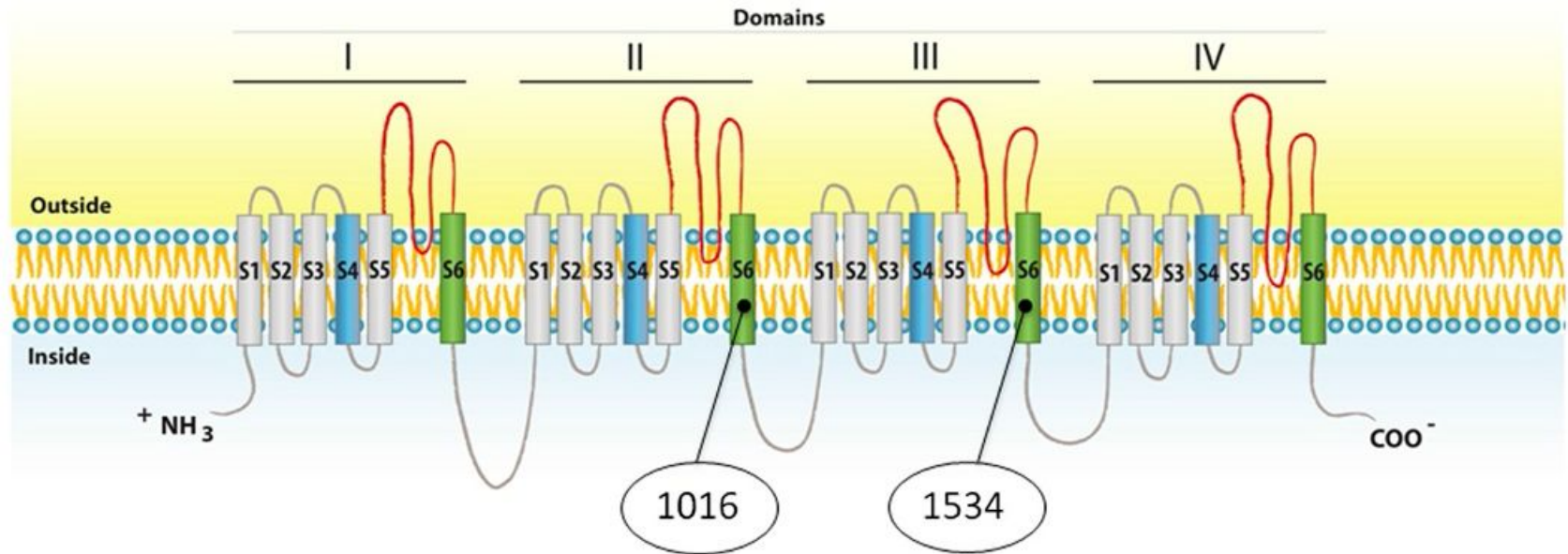
<https://www.ensembl.org/Help/Faq?id=335>

Query #3: What type of information is associated with the kdr variant?

The screenshot displays the Ensembl genome browser interface for the *Anopheles gambiae* (AgamP4) genome. The main focus is on the variant **tmp_2L_2422652_A_T snp**. A text box highlights the definition: "A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved." Below this, a table shows the variant's presence across transcripts, with a note that it overlaps 13 transcripts, has 2681 across genotypes, and is associated with 1 phenotype. The "Explore this variant" section includes five interactive cards: Genomic context, Genes and regulation, Flanking sequence (showing the sequence ATTGATT CCGCGCTG TCATGCT), Population genetics, and Phenotype data. A tooltip instruction reads: "Hover with the mouse in the items with dotted line for definitions." At the bottom, there are links for "Using the website" and "Reference materials".

Variant consequences

Kdr, 1014



Examples of other mutations in *Aedes aegypti*

Linss et al. 2014

Variant consequences

- Because missense (=nonsynonymous) variants ---> change in amino acid sequences
- In consequence, they may also disrupt or alter protein function
- The algorithm SIFT is run to score these changes, to predict if the protein function may change or not

Variant consequences

- SIFT calculates preservation of the aa sequence and domain in relationship to its orthologues.
- Predictions are NOT based on experimental evidence
- Interpretation:

Deleterious

Tolerated



Blue means low confidence values

Query #4: Find other mutations in AGAP004707 which also may be associated with pyrethroid resistance

The screenshot shows the VectorBase website interface. At the top, there is a navigation bar with links for 'About', 'Downloads', 'Tools', 'Data', 'Help', 'Community', 'Contact us', and 'More'. A search bar is located on the right. Below the navigation bar, the species 'Anopheles gambiae (AgamP4)' is identified, along with its location '2L:2,358,158-2,431,817' and a 'Gene page' button.

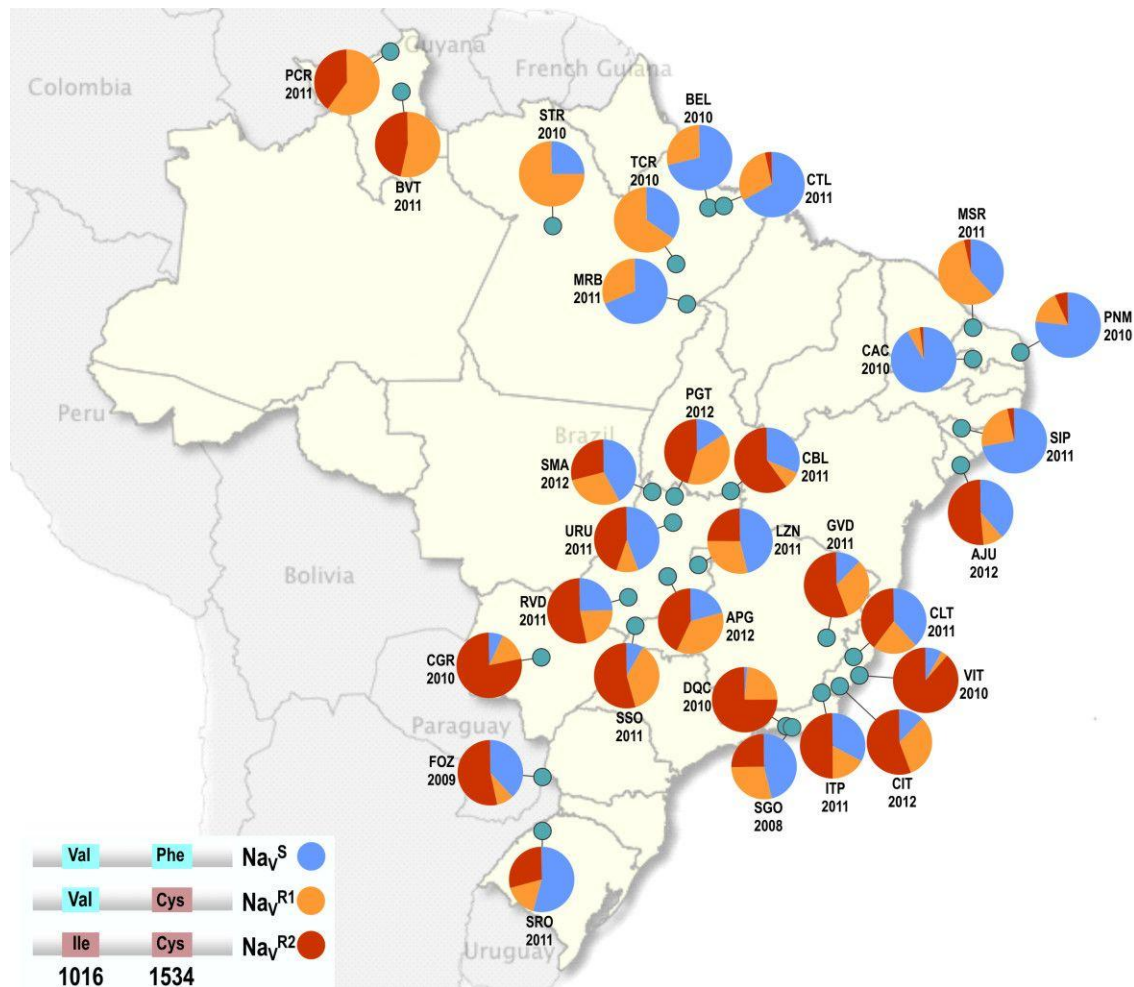
The main content area features a large grey box with the text: 'It shows the phenotypes, diseases and traits associated with this gene variant'. Below this, the 'Location' is specified as 'Chromosome 2L, 2,358,158-2,431,817 forward strand' and 'AgamP4:CM000356.1'. The 'About this gene' section states: 'This gene has 13 transcripts (splice variants), 46 orthologs, 10 paralogues and is associated with 1 phenotype'. A 'Show transcript table' button is present.

The 'Phenotypes' section is highlighted in green in the left sidebar. The main content area shows the heading 'Phenotype(s), disease(s) and trait(s) associated with this gene AGAP004707'. Below this, a table lists associated phenotypes:

Phenotype, disease and trait	Source	Study
resistance to treatment with the insecticide permethrin (BRC:1000014)	ABCMdb	PMID:21091761-d

Below the table, there are two lines of text: 'No phenotype, disease or trait has been associated with variants in this gene' and 'No phenotype, disease or trait associated with orthologues of this gene in other species'. At the bottom, there is a footer with 'VectorBase - View 1938 - June 2019 at 14:22:56' and 'World VectorBase | Contact Us | Help'.

Allele frequencies and geographic distribution



Kdr alleles in Brazilian *A. aegypti* populations
Linss et al. 2014

Query #5: What is the frequency of the mutant alleles?

The screenshot displays a genomic data browser interface. On the left, a navigation menu includes categories like 'Gene families', 'Orthologs', 'Phenotypes', and 'Genetic Variation', with 'Variant table' selected. The main content area shows a 'Variant table' for a specific gene. The table is filtered by 'SIFT: All' and 'Consequence: missense variant'. A dropdown menu is open over the 'AA ind' column, with 'AA view' highlighted. The table contains the following data:

Variant ID	Chr:bp	Allele	Class	Score	Conseq. Type	AA	AA ind	SIFT	Transcript
trp_2L_226020 T.G.A	2L:2260204	G/A	SNP	VDPC 103	missense variant	D/N	33	0.06	AGA*004702- RA
trp_2L_226020 Z.A.I	2L:2260200	A/T	SNP	VSPC 103	missense variant	D/V	54	0.01	AGA*004702- RA
trp_2L_226020 Z.G.T	2L:2260210	G/T	SNP	VDPC 103	missense variant	G/C	80	0	AGA*004702- RA
trp_2L_226020 Z.C.T	2L:2260200	C/T	SNP	VDP0000- 103	missense variant	P/L	81	0.1	AGA*004702- RA
trp_2L_226018 E.A.G	2L:2260188	A/G	SNP	VSP0000- 103	missense variant	K/R	257	0.12	AGA*004702- RA
trp_2L_226017 T.G.A	2L:2260177	G/A	SNP	VDP0000- 103	missense variant	R/K	200	0.25	AGA*004702- RA
trp_2L_226020 Z.A.A	2L:2260211	G/A	SNP	VDP0000- 103	missense variant	E/R	325	0.26	AGA*004702- RA

Query #6: What is the geographic distribution
of the mutant alleles?

Demo

(instructions in the speaker notes)

VEP, the variant effect predictor

Predicts the **effect of variants** (SNPs, insertions, deletions, CNVs, or structural variants):



Input format

- Genomic coordinates
- variant call format, VCF
- Variant IDs

Results:

Affected gene, transcript, and protein sequence

SIFT scores

Frequency data

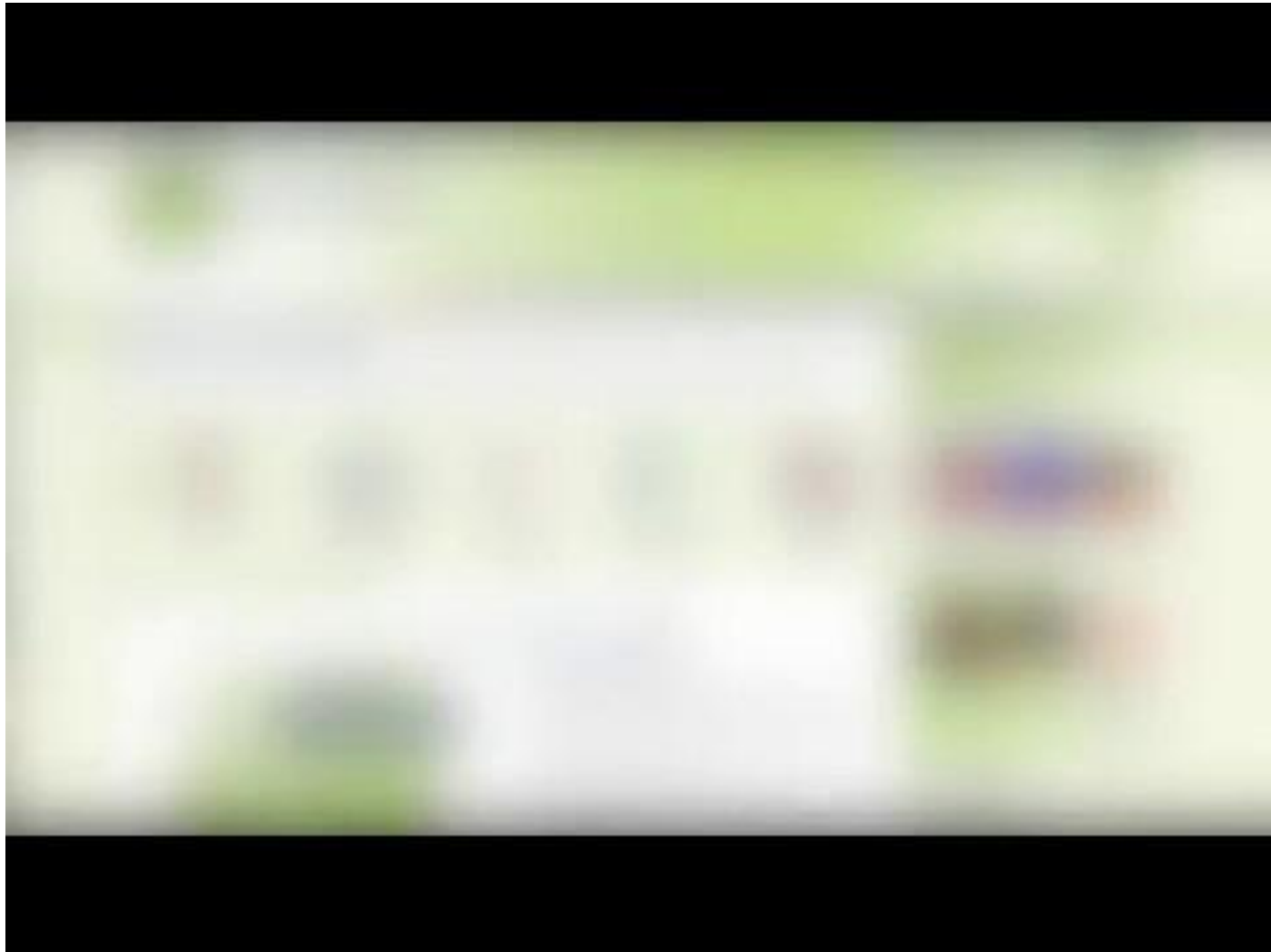
Regulatory consequences

Splicing consequences

Literature citations

Query #7: What are the variant effects of my own data?

Are my variants already annotated in VectorBase?, Which genes are affected by my variants?, My variants affect gene regulation?



VEP for non-VectorBase species

- Send us an email
- We may be able to help you run your own local instance of the tool

Summary: Variation Data

How to browse for variant data? Use the Genome Browser different tabs

- Gene
- Transcript
- Location
- Variant



The screenshot shows the VectorBase website interface. At the top left is the VectorBase logo, which includes a circular image of a mosquito and the text "VectorBase Bioinformatics Resource for Invertebrate Vectors of Human Pathogens". To the right of the logo is a navigation menu with links for "About", "Downloads", "Tools", "Data", "Help", and "Community". Below the logo is a dropdown menu for "Anopheles gambiae (AgamP4)". At the bottom of the interface, there are four tabs: "Location: 2L:2,350,505-2,440,582", "Gene: para", "Transcript: AGAP004707-RA", and "Variant: tmp_2L_2422652_A_T". The "Variant" tab is currently selected and highlighted in blue.

How to search for more information or help?

E-mail us at
info@vectorbase.org

Thank you!