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 <u>your own</u> 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
Aedes aegypti Liverpool	AaegL3	332,445	4,081
Aedes aegypti LVP_AGWG	AaegL5	313,612	1,761
Anopheles arabiensis Dongola	AaraD1	10,164,339	1,022,752
Anopheles culicifacies A-37	AculA1	9,154,354	900,063
Anopheles epiroticus Epiroticus2	AepiE1	3,281,528	267,653
Anopheles farauti FAR1	AfarF2	7,226,209	871,433
Anopheles funestus FUMOZ	AfunF1	12,920,105	987,886
Anopheles gambiae PEST	AgamP4	59,236,057	2,168,425
Anopheles melas CM1001059_A	AmelC2	3,677,409	418,730
Anopheles merus MAF	AmerM2	6,053,485	558,680
Anopheles minimus MINIMUS1	AminM1	4,206,083	222,877
Anopheles quadriannulatus SANGWE	AquaS1	10,137,655	931,982
Anopheles sinensis SINENSIS	AsinS2	5,841,551	409,698
Anopheles stephensi SDA-500	AsteS1	5,856,684	574,778
Anopheles stephensi Indian	AsteI2	366,367	0
Biomphalaria glabrata BB02	BglaB1	10,031,395	0
Culex quinquefasciatus Johannesburg	CpipJ2	2	0
Ixodes scapularis 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

https://www.vectorbase.org/release/release-vb-2018-06

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

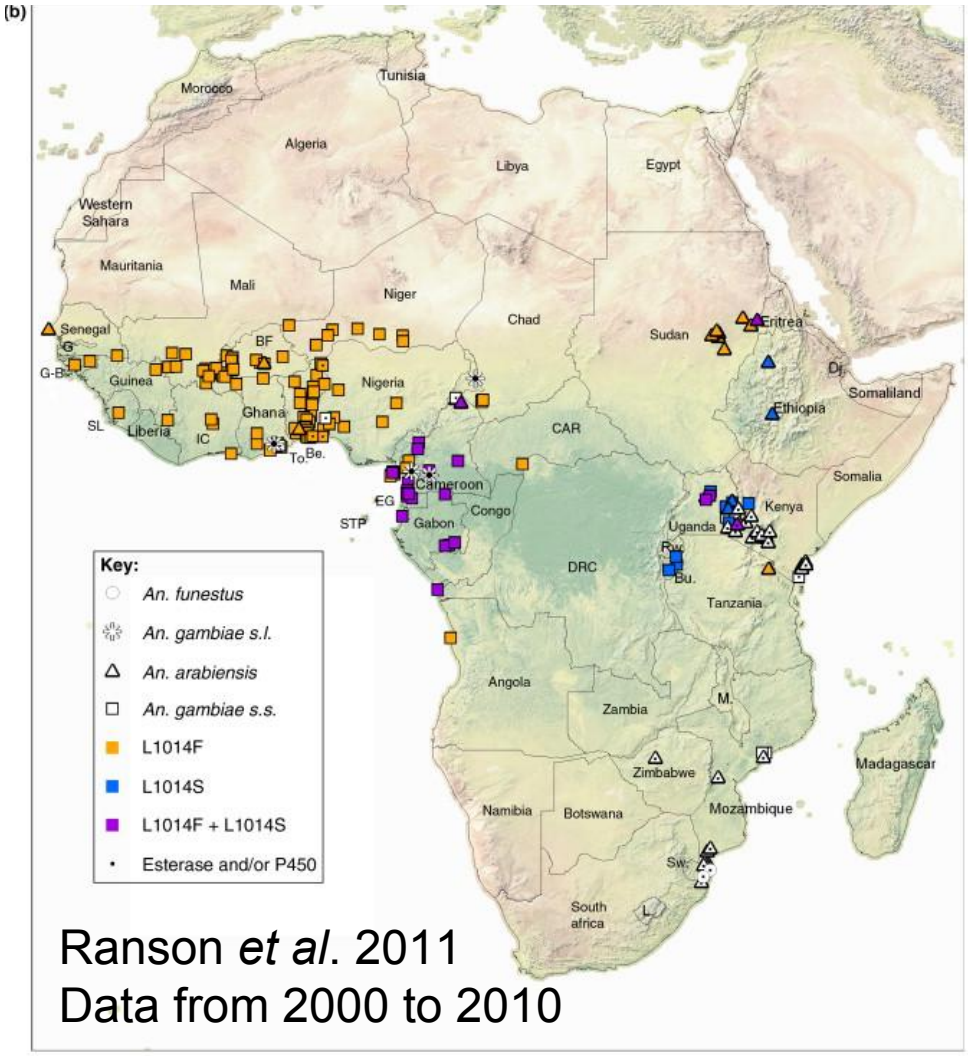
G is the allele in the contig used in the reference genome

- 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!

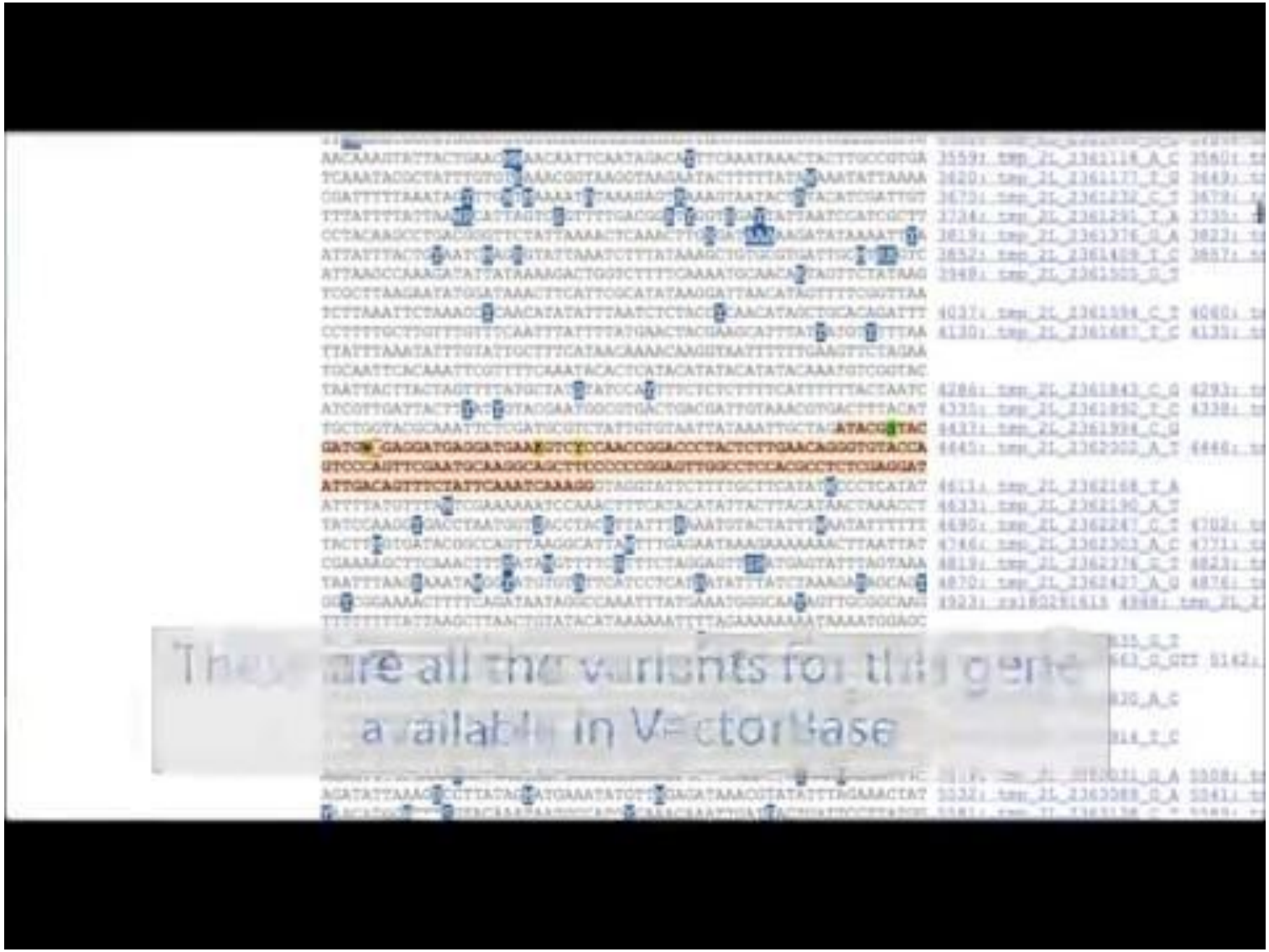
Reference and Alternative alleles: Sample case

- *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)

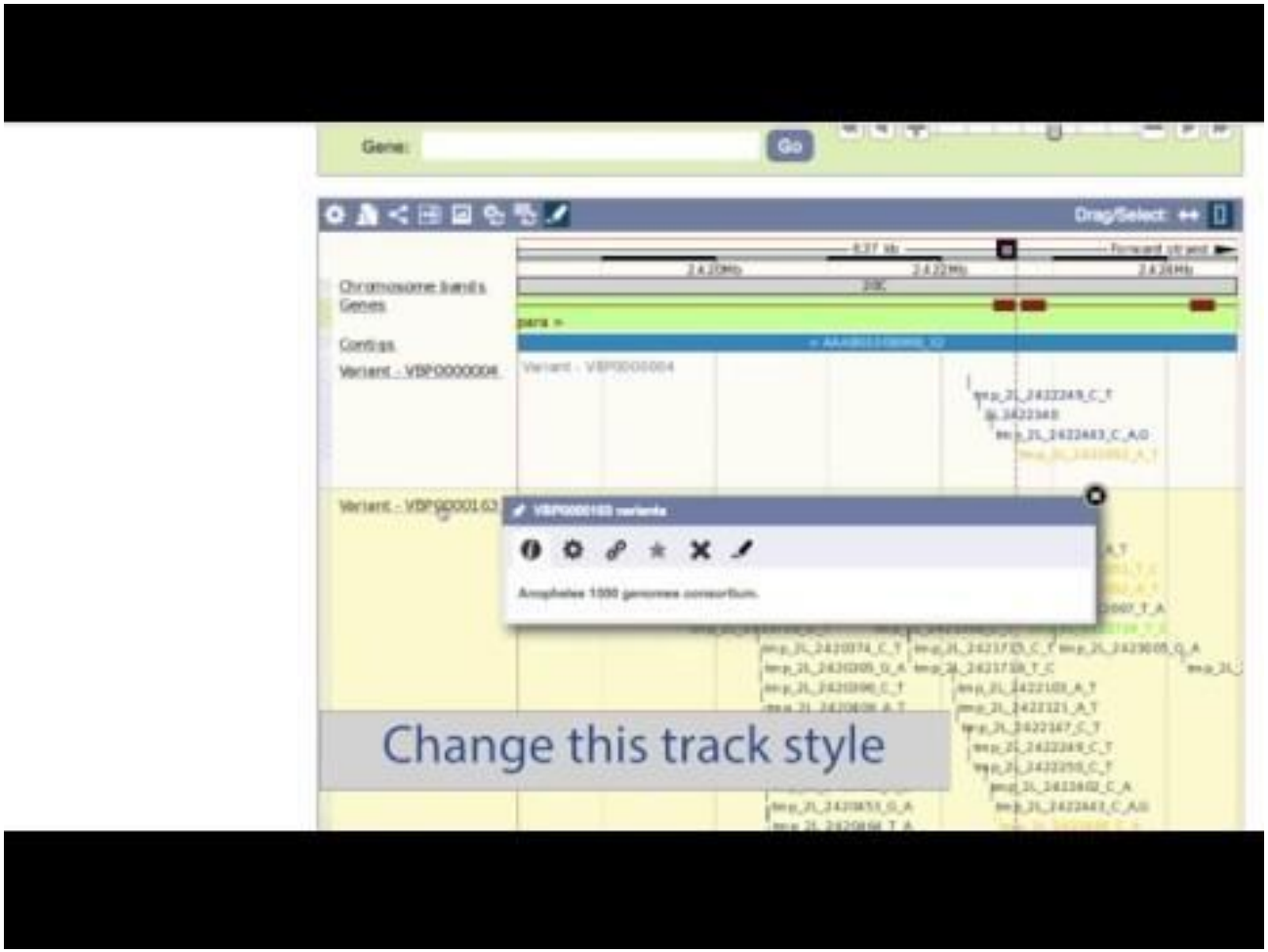


Ranson *et al*. 2011
Data from 2000 to 2010

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



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

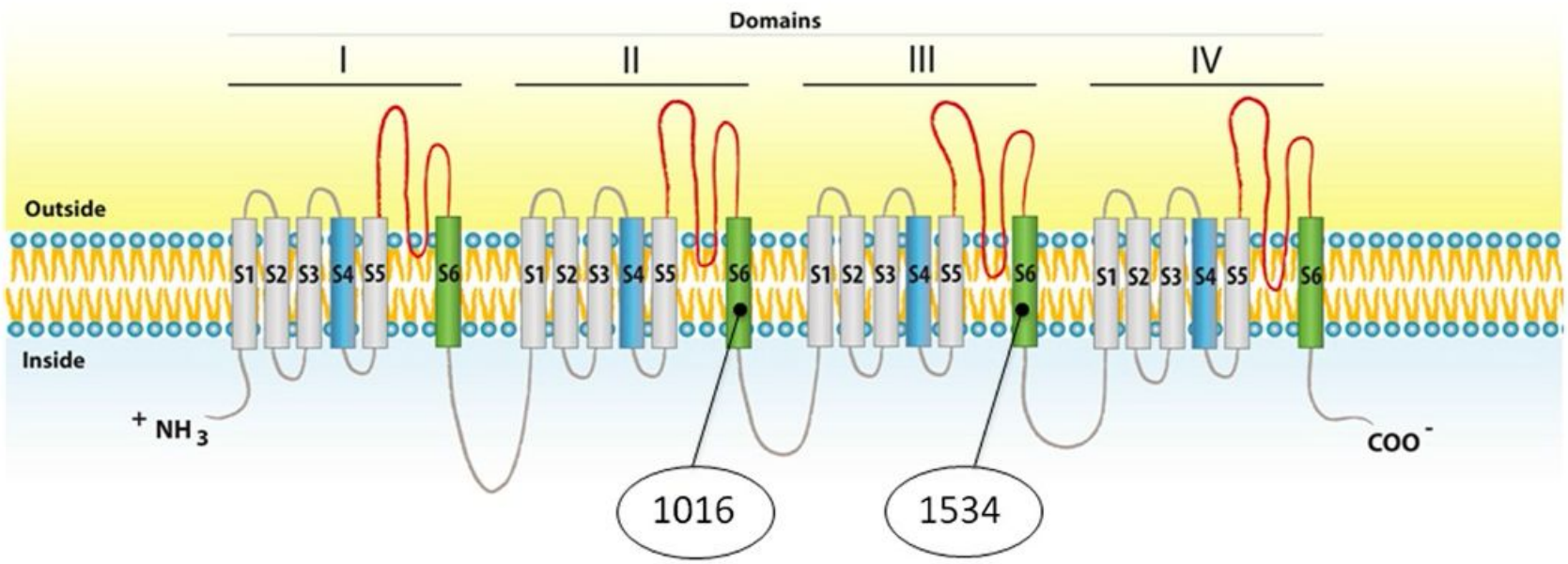


Track styles:
https://www.ensembl.org/Help/Faq?id=335

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

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:

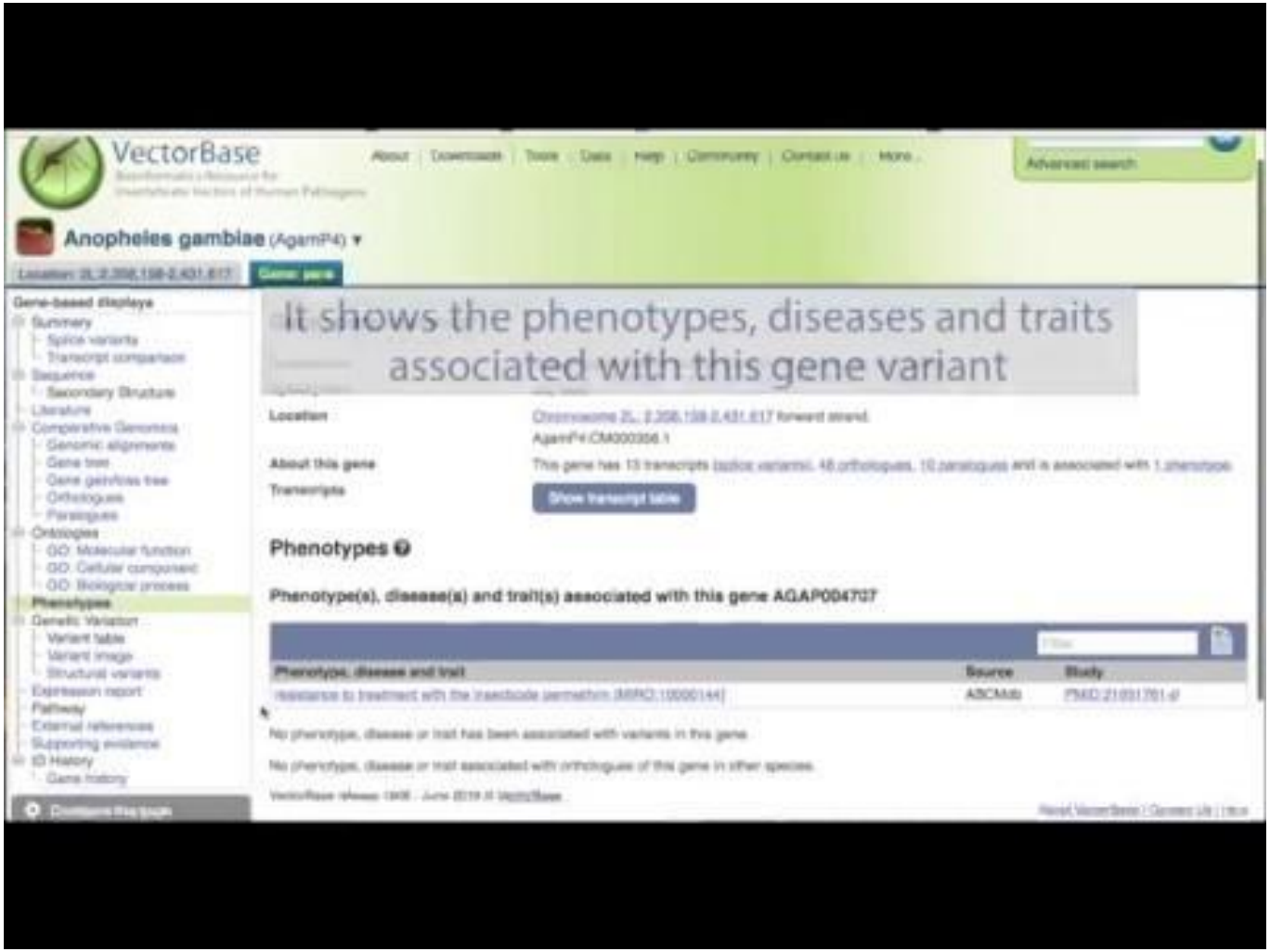
Deleterious Tolerated

0 ---|---------- 1

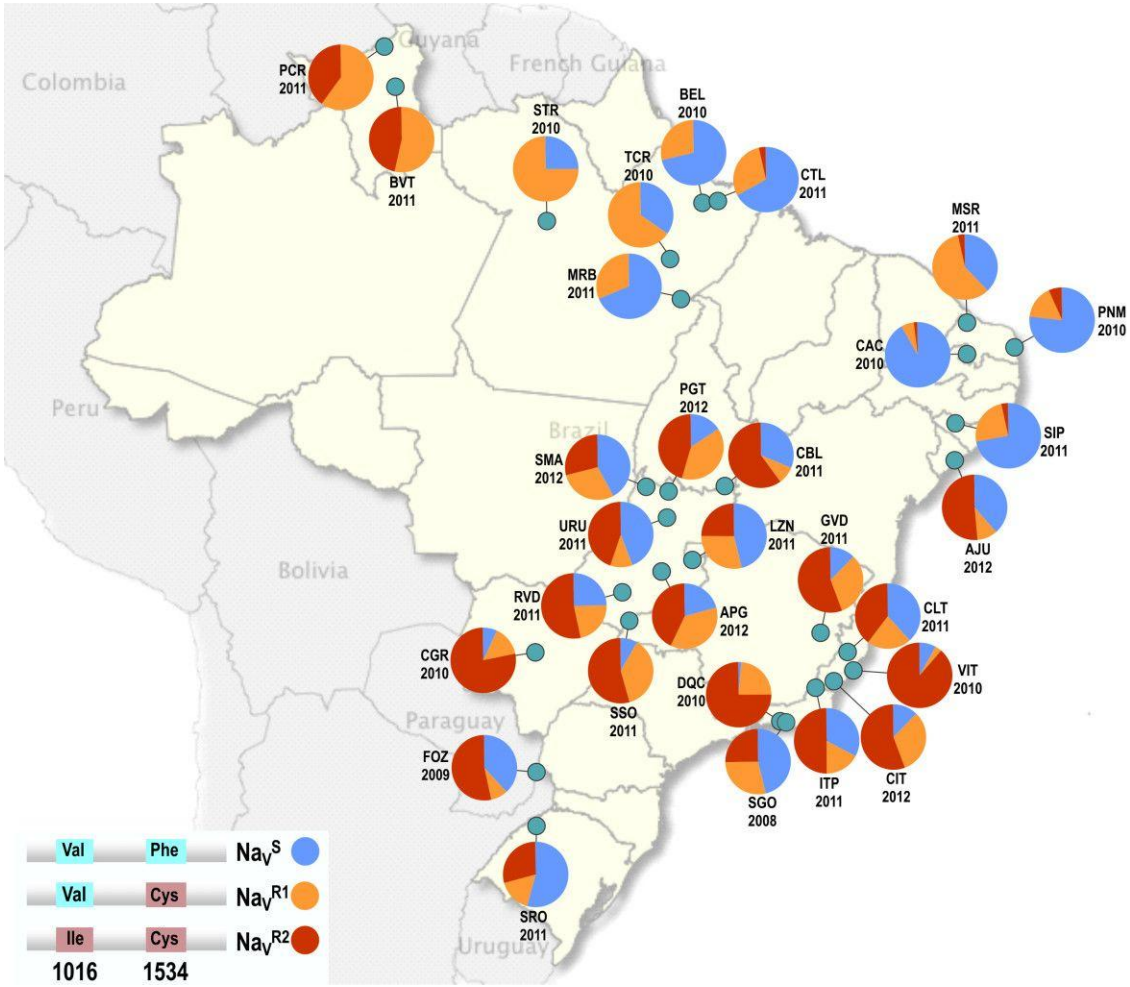
0.05

Blue means low confidence values

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

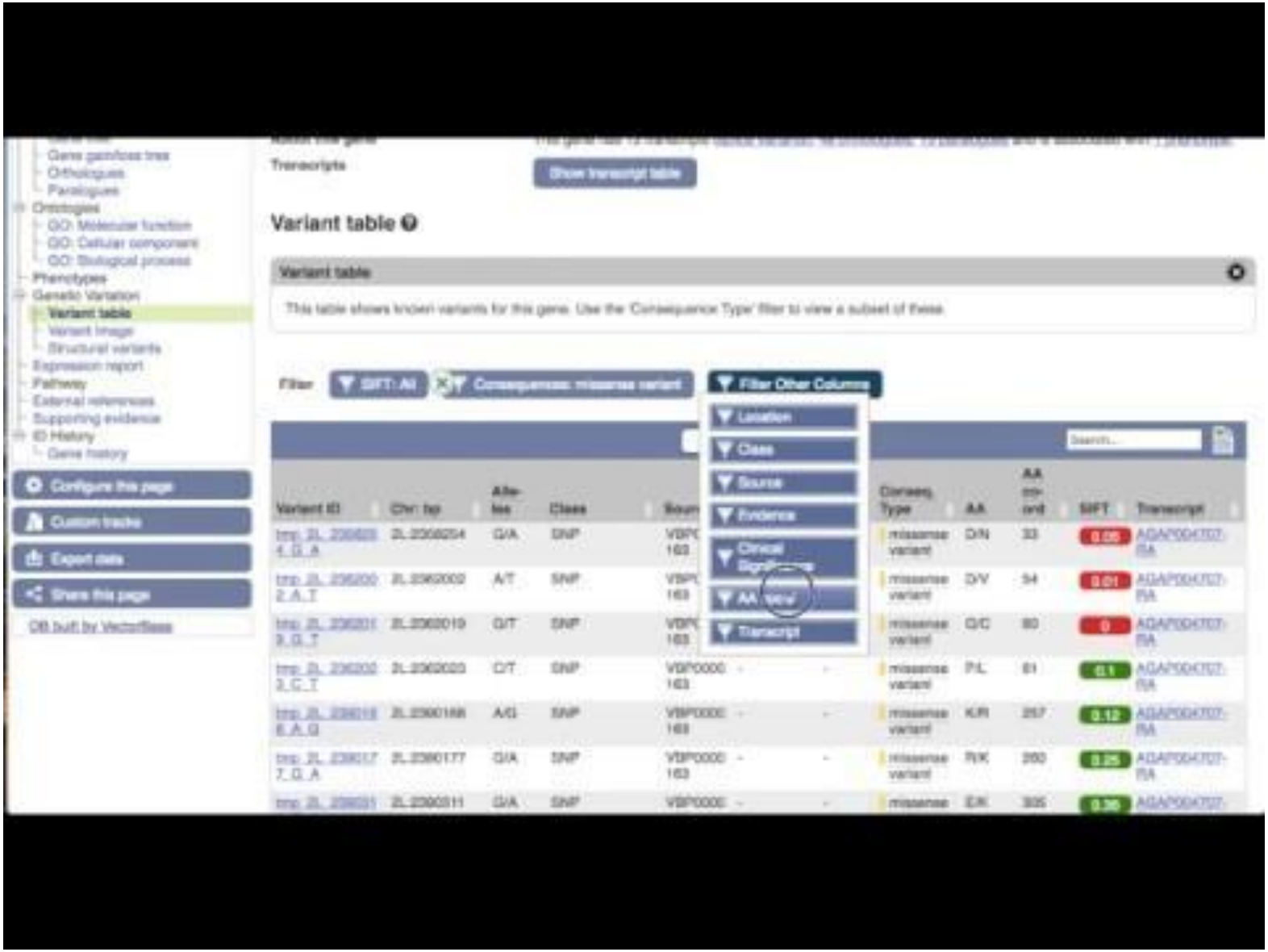


Allele frequencies and geographic distribution



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

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



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 <u>Genome Browser</u> different tabs

- Gene
- Transcript
- Location
- Variant

How to search for more information or help?

E-mail us at

info@vectorbase.org

Thank you!