Non-model arthropod assembly: past, present and future

Scott Emrich
On behalf of VectorBase and i5K

Eck Institute for Global Health
University of Notre Dame

VectorBase
Bioinformatics Resource for Invertebrate Vectors of Human Pathogens

Brief overview

- Brief self introduction
- The VectorBase BRC
- “10 simple rules for a successful genome project” – ode to PlOS series

DIRTY ROTTEN BUGS?

Arthropods Unite to Tell Their Side of the Story

Written and illustrated by GILLES BONOTAUX

DON'T LET
THIS BAD BUG
BITE YOU

A brief introduction to BRCs

VectorBase
Bioinformatics Resource for
Invertebrate Vectors of Human Pathogens

EuPathDB
Eukaryotic Pathogen Database Resources
http://EuPathDB.org

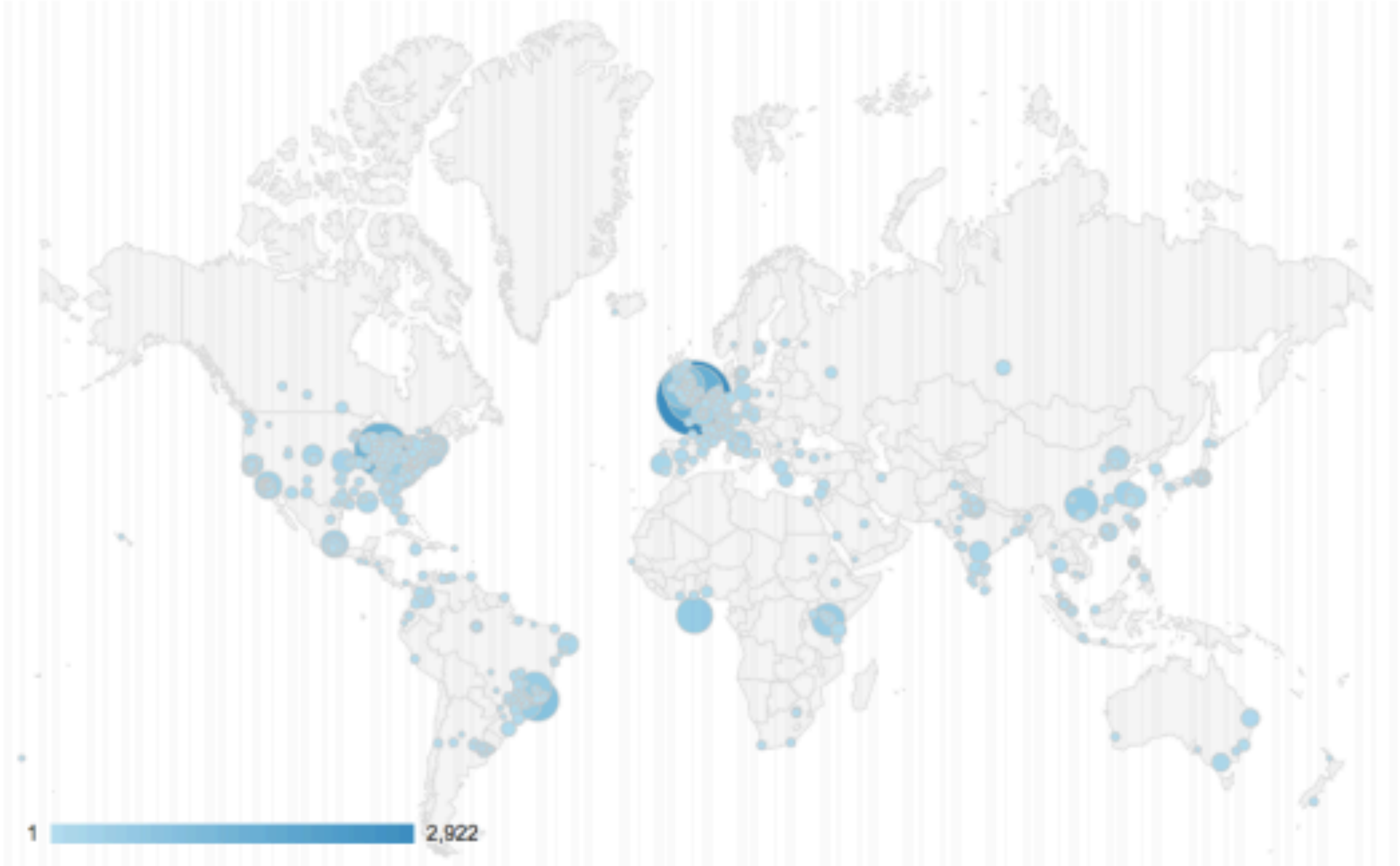
IRD Influenza Research Database

PATRIC
PATHOSYSTEMS RESOURCE INTEGRATION CENTER

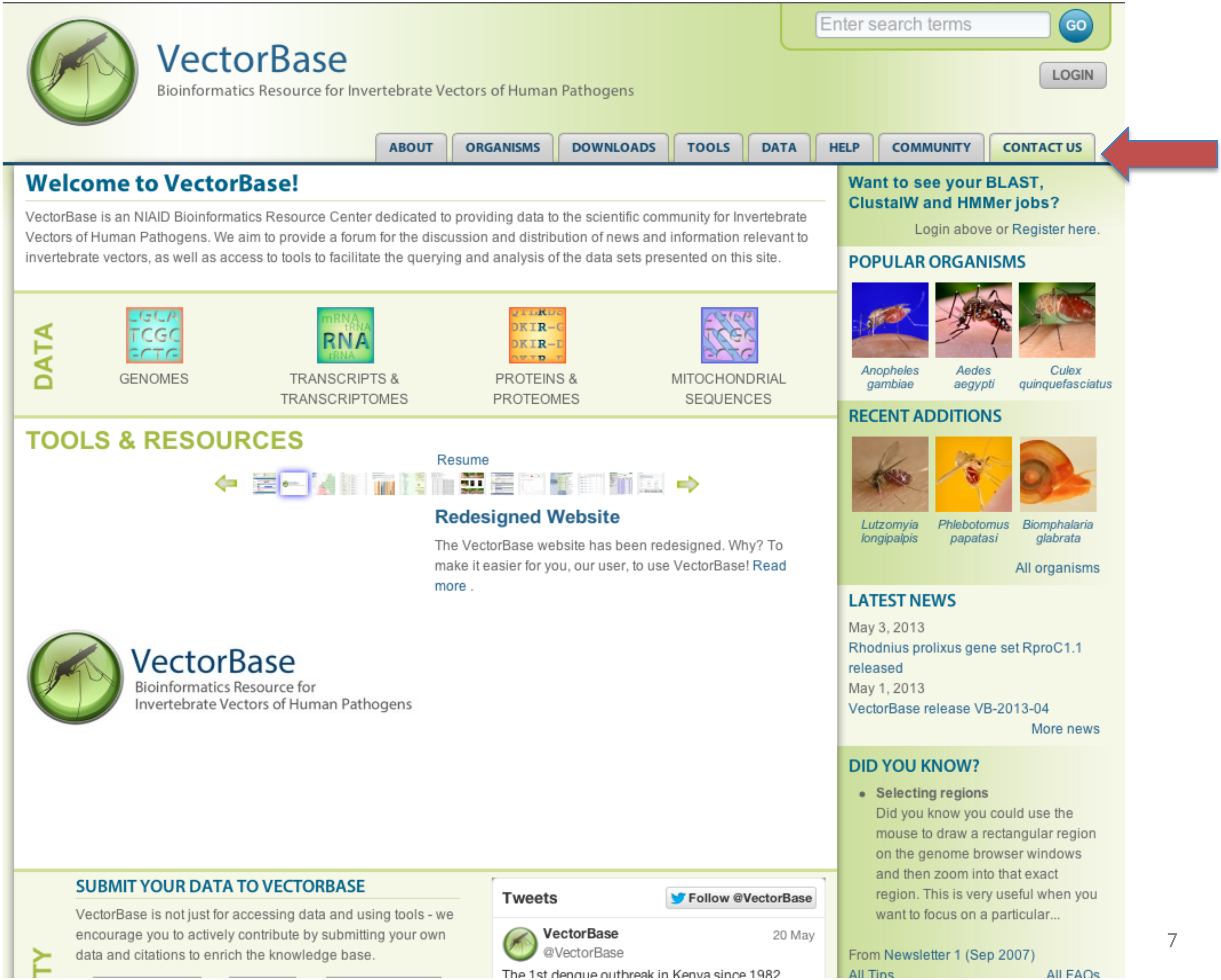
ViPR
Virus Pathogen Resource

- VectorBase is a genome resource for invertebrate vectors of human pathogens
- Funded by NIH-NIAID as part of a wider group of NIAID BRCs (see above) for biodefense and emerging and re-emerging infectious diseases
- Third contract started Fall 2014 (for up to 5 more years)

VectorBase Scientific User Community by Loci



VectorBase Scientific User Community by Loci



VectorBase

Bioinformatics Resource for Invertebrate Vectors of Human Pathogens

Welcome to VectorBase!

VectorBase is an NIAID Bioinformatics Resource Center dedicated to providing data to the scientific community for Invertebrate Vectors of Human Pathogens. We aim to provide a forum for the discussion and distribution of news and information relevant to invertebrate vectors, as well as access to tools to facilitate the querying and analysis of the data sets presented on this site.

TOOLS & RESOURCES

Redesigned Website

The VectorBase website has been redesigned. Why? To make it easier for you, our user, to use VectorBase! Read more .

VectorBase

Bioinformatics Resource for Invertebrate Vectors of Human Pathogens

SUBMIT YOUR DATA TO VECTORBASE

VectorBase is not just for accessing data and using tools - we encourage you to actively contribute by submitting your own data and citations to enrich the knowledge base.

Tweets

VectorBase 20 May
@VectorBase
The 1st dengue outbreak in Kenya since 1982

Want to see your BLAST, ClustalW and HMMer jobs?

Login above or Register here.

POPULAR ORGANISMS

Anopheles gambiae *Aedes aegypti* *Culex quinquefasciatus*

RECENT ADDITIONS

Lutzomyia longipalpis *Phlebotomus papatasi* *Biomphalaria glabrata*

All organisms

LATEST NEWS

May 3, 2013
Rhodnius prolixus gene set RproC1.1 released
May 1, 2013
VectorBase release VB-2013-04

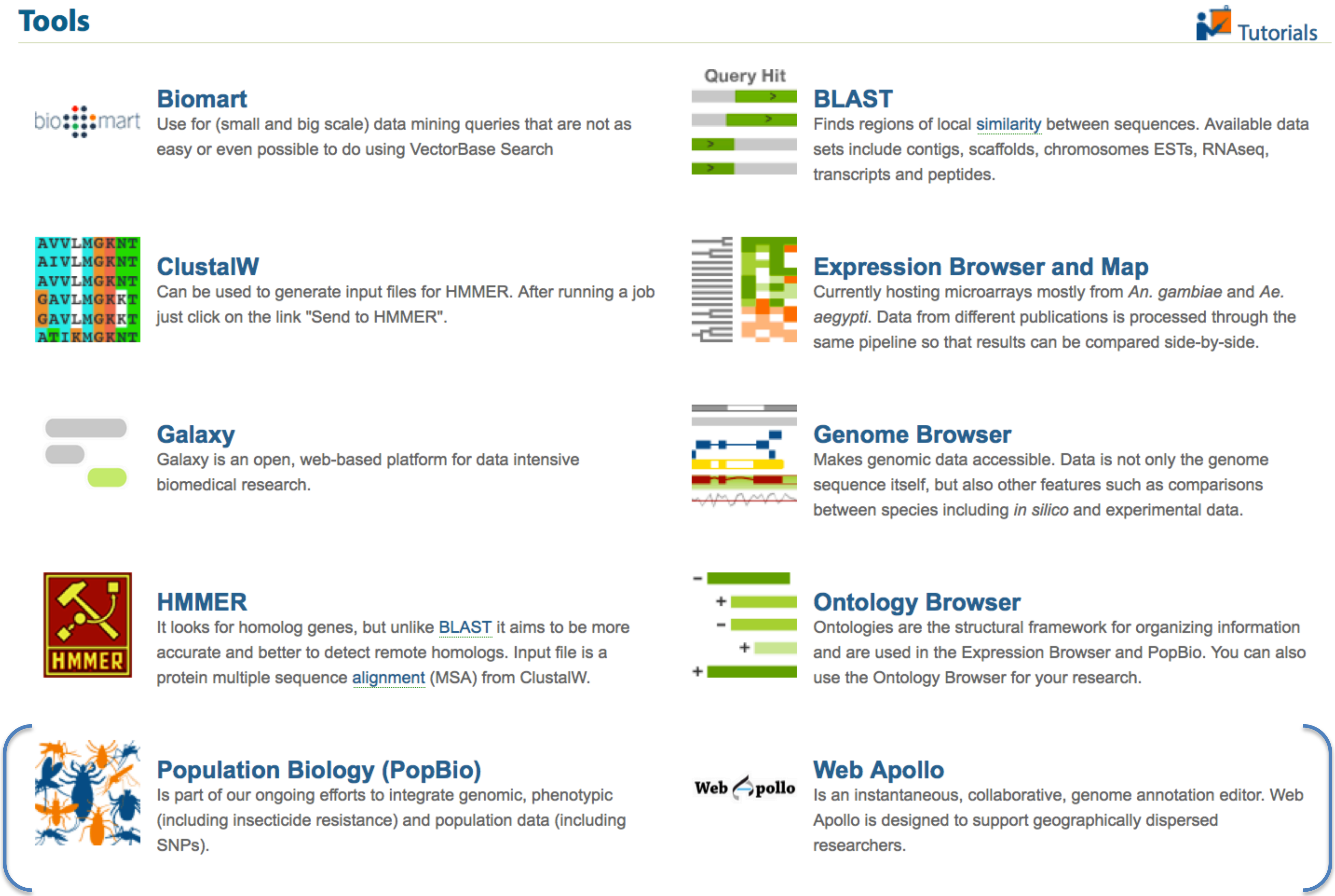
More news

DID YOU KNOW?

- **Selecting regions**
  Did you know you could use the mouse to draw a rectangular region on the genome browser windows and then zoom into that exact region. This is very useful when you want to focus on a particular...

From Newsletter 1 (Sep 2007)

All Tips All FAQs



Tools

Tutorials

Biomart

Use for (small and big scale) data mining queries that are not as easy or even possible to do using VectorBase Search

ClustalW

Can be used to generate input files for HMMER. After running a job just click on the link "Send to HMMER".

Galaxy

Galaxy is an open, web-based platform for data intensive biomedical research.

HMMER

It looks for homolog genes, but unlike BLAST it aims to be more accurate and better to detect remote homologs. Input file is a protein multiple sequence alignment (MSA) from ClustalW.

Population Biology (PopBio)

Is part of our ongoing efforts to integrate genomic, phenotypic (including insecticide resistance) and population data (including SNPs).

BLAST

Finds regions of local similarity between sequences. Available data sets include contigs, scaffolds, chromosomes ESTs, RNAseq, transcripts and peptides.

Expression Browser and Map

Currently hosting microarrays mostly from *An. gambiae* and *Ae. aegypti*. Data from different publications is processed through the same pipeline so that results can be compared side-by-side.

Genome Browser

Makes genomic data accessible. Data is not only the genome sequence itself, but also other features such as comparisons between species including *in silico* and experimental data.

Ontology Browser

Ontologies are the structural framework for organizing information and are used in the Expression Browser and PopBio. You can also use the Ontology Browser for your research.

Web Apollo

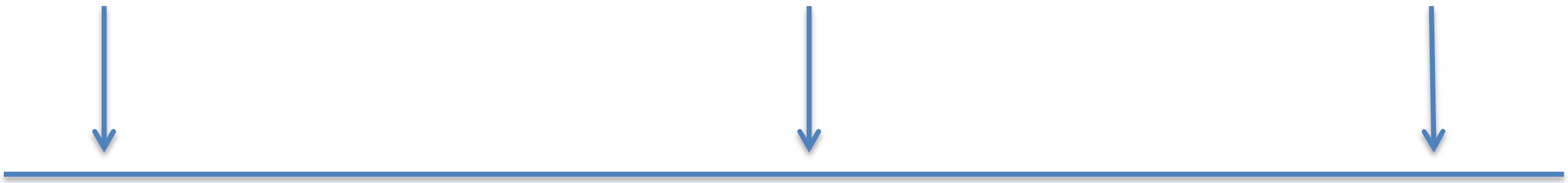
Is an instantaneous, collaborative, genome annotation editor. Web Apollo is designed to support geographically dispersed researchers.

Community gene annotation using WebApollo

Pre Web
Apollo

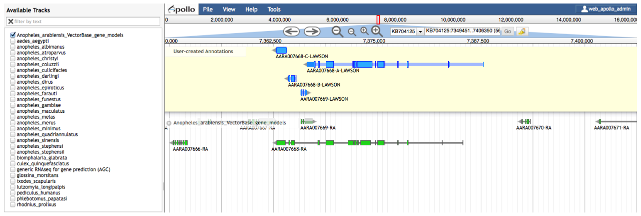
Early 2014
Web Apollo 1

December 2015
Web Apollo 2



User Downloads and edits GFF with text editor or WebApollo.

Inconsistent GFF with many errors.



☑ Anopheles_arabiensis_VectorBase_gene_models
☐ aedes_aegypti
☐ anopheles_albimanus
☐ anopheles_atroparvus
☐ anopheles_christyi
☐ anopheles_coluzzii
☐ anopheles_culicifacies
☐ anopheles_darlingi
☐ anopheles_dirus
☐ anopheles_epiroticus
☐ anopheles_farauti
☐ anopheles_funestus
☐ anopheles_gambiae
☐ anopheles_maculatus
☐ anopheles_melas
☐ anopheles_merus
☐ anopheles_minimus
☐ anopheles_quadriannulatus
☐ anopheles_sinensis
☐ anopheles_stephensi
☐ anopheles_stephensil
☐ biomphalaria_glabrata
☐ culex_quinquefasciatus
☐ generic RNAseq for gene prediction (AGC)
☐ glossina_morsitans
☐ ixodes_scapularis
☐ lutzomyia_longipalpis
☐ pediculus_humanus
☐ phlebotomus_papatasi
☐ rhodnius_prolixus

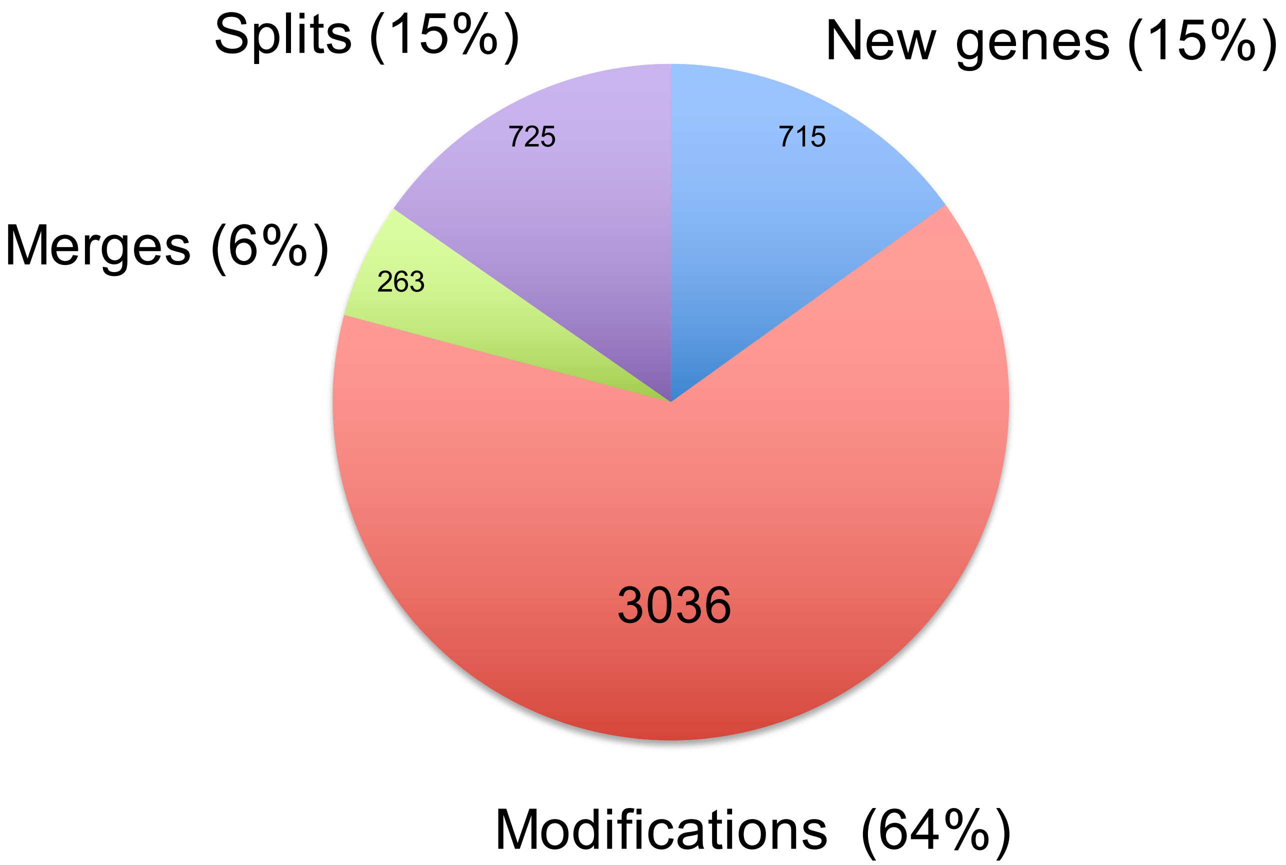
Common application for all users. ✓

Need one instance per species.
Hard to access meta data.

One Instance for all species ✓
Reports on demand ✓

36 species
>100 users

Summary of Community Annotation

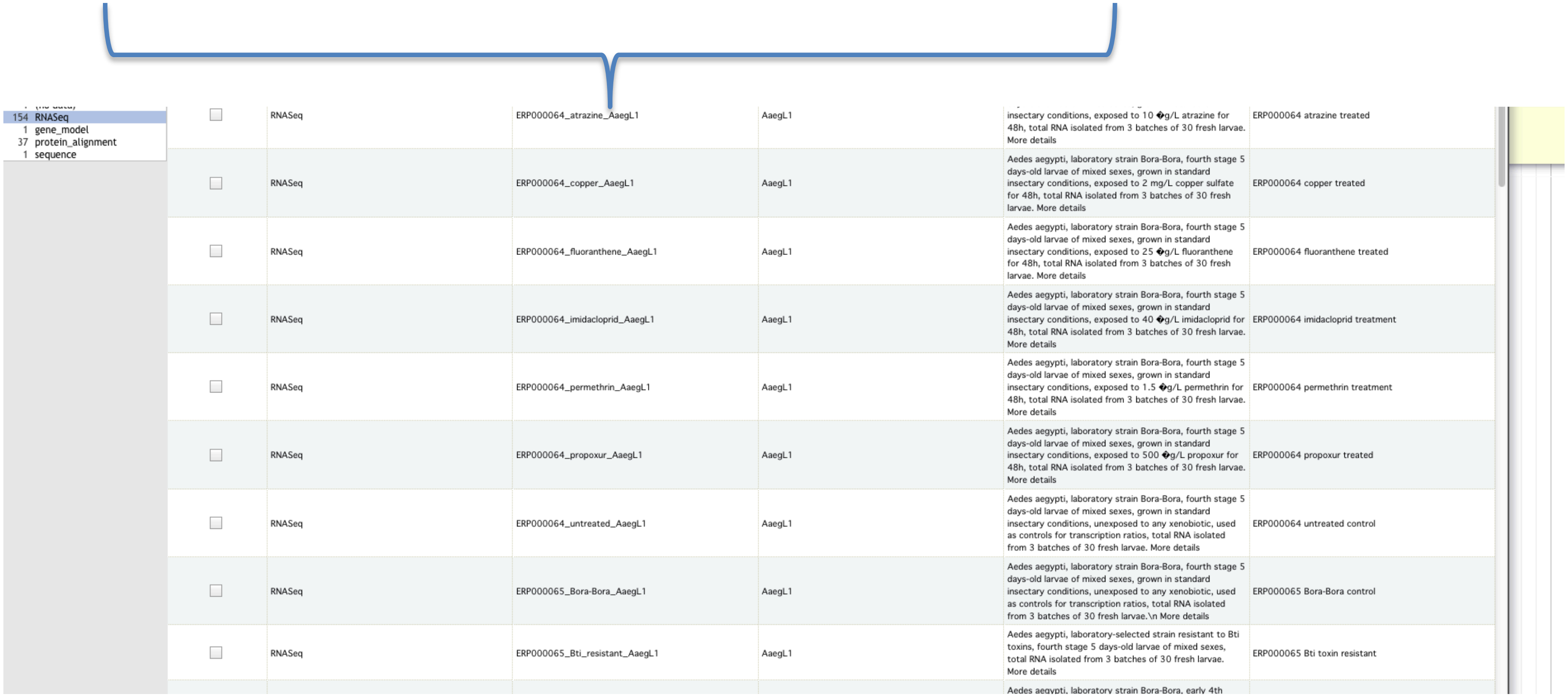


Total Events 4739
29 Species

- 1/3 of annotations are major gene-model changes (splits, merges)
- 2/3 of annotations are minor modifications
- 2/3 of modifications are to the CDS
- ½ of CDS modifications gain sequence
- Accrued relevant InterPro Domains.

Searchable track Meta Data

Publications ID, External Database ID, Description ….

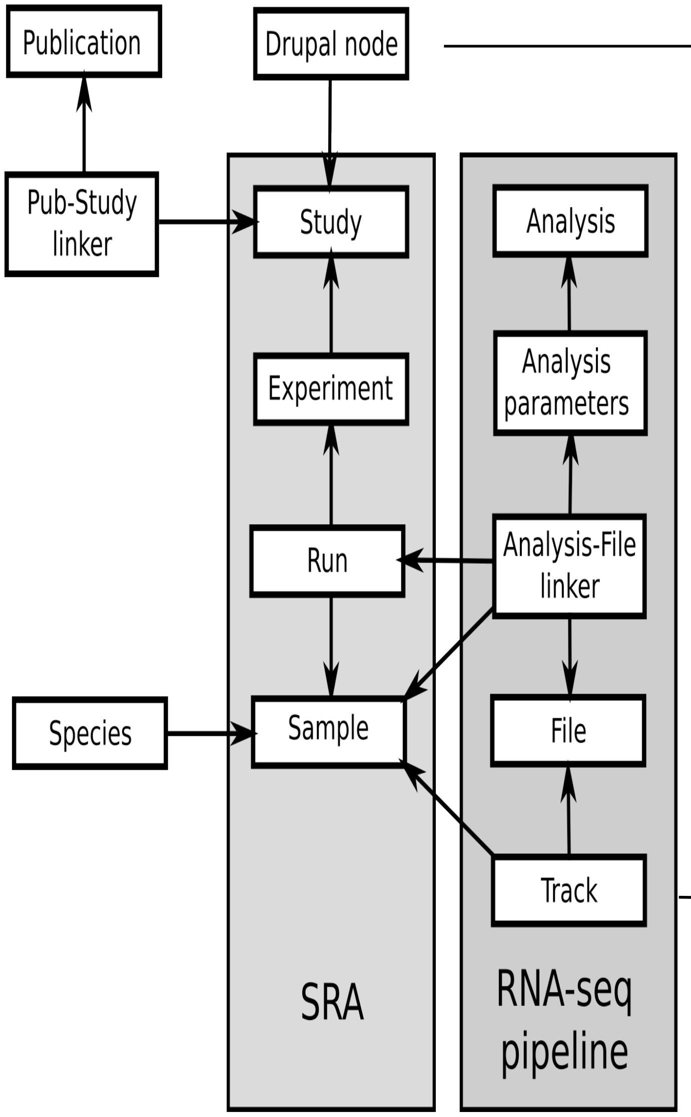


(no data)
154 RNASeq
1 gene_model
37 protein_alignment
1 sequence

□	RNASeq	ERP000064_atrazine_AaegL1	AaegL1	insectary conditions, exposed to 10 �g/L atrazine for 48h, total RNA isolated from 3 batches of 30 fresh larvae. More details	ERP000064 atrazine treated
□	RNASeq	ERP000064_copper_AaegL1	AaegL1	Aedes aegypti, laboratory strain Bora-Bora, fourth stage 5 days-old larvae of mixed sexes, grown in standard insectary conditions, exposed to 2 mg/L copper sulfate for 48h, total RNA isolated from 3 batches of 30 fresh larvae. More details	ERP000064 copper treated
□	RNASeq	ERP000064_fluoranthene_AaegL1	AaegL1	Aedes aegypti, laboratory strain Bora-Bora, fourth stage 5 days-old larvae of mixed sexes, grown in standard insectary conditions, exposed to 25 �g/L fluoranthene for 48h, total RNA isolated from 3 batches of 30 fresh larvae. More details	ERP000064 fluoranthene treated
□	RNASeq	ERP000064_imidacloprid_AaegL1	AaegL1	Aedes aegypti, laboratory strain Bora-Bora, fourth stage 5 days-old larvae of mixed sexes, grown in standard insectary conditions, exposed to 40 �g/L imidacloprid for 48h, total RNA isolated from 3 batches of 30 fresh larvae. More details	ERP000064 imidacloprid treatment
□	RNASeq	ERP000064_permethrin_AaegL1	AaegL1	Aedes aegypti, laboratory strain Bora-Bora, fourth stage 5 days-old larvae of mixed sexes, grown in standard insectary conditions, exposed to 1.5 �g/L permethrin for 48h, total RNA isolated from 3 batches of 30 fresh larvae. More details	ERP000064 permethrin treatment
□	RNASeq	ERP000064_propoxur_AaegL1	AaegL1	Aedes aegypti, laboratory strain Bora-Bora, fourth stage 5 days-old larvae of mixed sexes, grown in standard insectary conditions, exposed to 500 �g/L propoxur for 48h, total RNA isolated from 3 batches of 30 fresh larvae. More details	ERP000064 propoxur treated
□	RNASeq	ERP000064_untreated_AaegL1	AaegL1	Aedes aegypti, laboratory strain Bora-Bora, fourth stage 5 days-old larvae of mixed sexes, grown in standard insectary conditions, unexposed to any xenobiotic, used as controls for transcription ratios, total RNA isolated from 3 batches of 30 fresh larvae. More details	ERP000064 untreated control
□	RNASeq	ERP000065_Bora-Bora_AaegL1	AaegL1	Aedes aegypti, laboratory strain Bora-Bora, fourth stage 5 days-old larvae of mixed sexes, grown in standard insectary conditions, unexposed to any xenobiotic, used as controls for transcription ratios, total RNA isolated from 3 batches of 30 fresh larvae.\n More details	ERP000065 Bora-Bora control
□	RNASeq	ERP000065_Bti_resistant_AaegL1	AaegL1	Aedes aegypti, laboratory-selected strain resistant to Bti toxins, fourth stage 5 days-old larvae of mixed sexes, total RNA isolated from 3 batches of 30 fresh larvae. More details	ERP000065 Bti toxin resistant
				Aedes aegypti, laboratory strain Bora-Bora, early 4th	

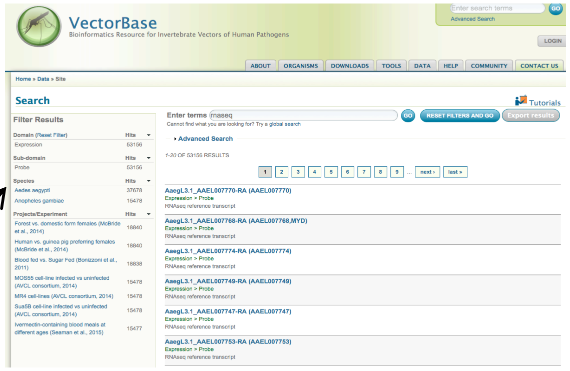
Several hundred RNASeq tracks with associated metadata have been automatically added to VectorBase

RNAseq tracking and analysis



Apache Solr

UCSC track hubs



VectorBase
Bioinformatics Resource for Invertebrate Vectors of Human Pathogens

Enter search terms GO
Advanced Search
LOGIN

ABOUT | ORGANISMS | DOWNLOADS | TOOLS | DATA | HELP | COMMUNITY | CONTACT US

Home » Data » Site

Search

Tutorials

Filter Results

Domain (Reset Filter)	Hits
Expression	53156

Sub-domain	Hits
Probe	53156

Species	Hits
Aedes aegypti	37678
Anopheles gambiae	15478

Projects/Experiment	Hits
Forest vs. domestic form females (McBride et al., 2014)	18840
Human vs. guinea pig preferring females (McBride et al., 2014)	18840
Blood fed vs. Sugar Fed (Bonizzoni et al., 2011)	18838
MOS55 cell-line infected vs uninfected (AVCL consortium, 2014)	15478
MR4 cell-lines (AVCL consortium, 2014)	15478
Sua5B cell-line infected vs uninfected (AVCL consortium, 2014)	15478
Ivermectin-containing blood meals at different ages (Seaman et al., 2015)	15477

Enter terms rnaseq GO RESET FILTERS AND GO Export results
Cannot find what you are looking for? Try a global search

Advanced Search

1-20 OF 53156 RESULTS

1 2 3 4 5 6 7 8 9 … next › last »

AaegL3.1_AAEL007770-RA (AAEL007770)
Expression > Probe
RNAseq reference transcript

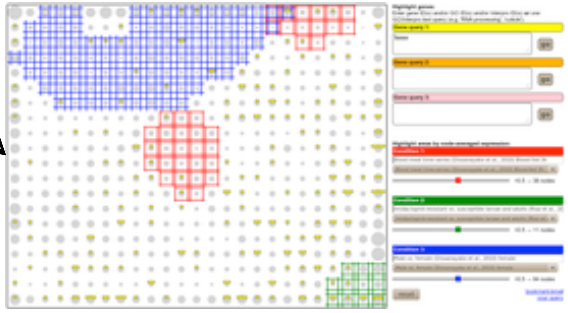
AaegL3.1_AAEL007768-RA (AAEL007768,MYD)
Expression > Probe
RNAseq reference transcript

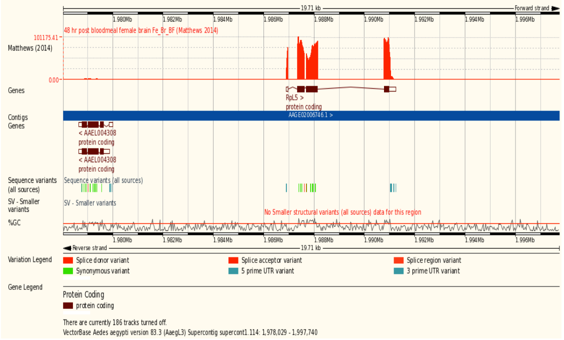
AaegL3.1_AAEL007774-RA (AAEL007774)
Expression > Probe
RNAseq reference transcript

AaegL3.1_AAEL007749-RA (AAEL007749)
Expression > Probe
RNAseq reference transcript

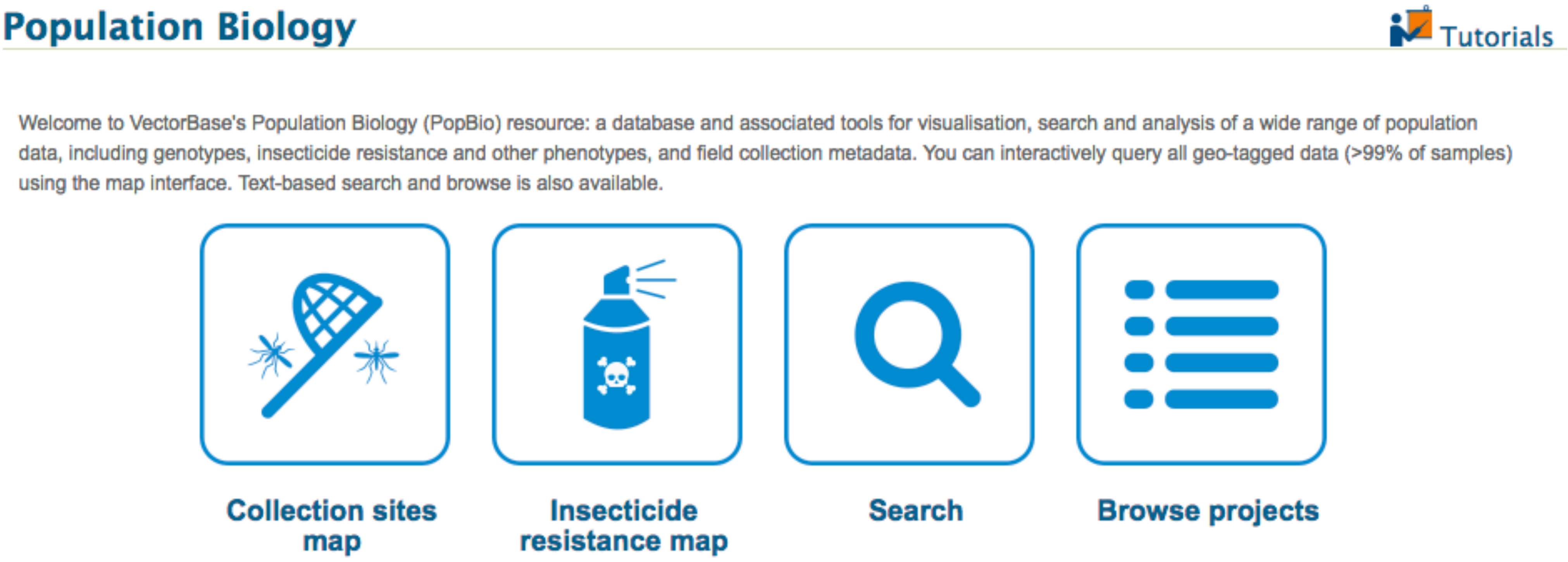
AaegL3.1_AAEL007747-RA (AAEL007747)
Expression > Probe
RNAseq reference transcript

AaegL3.1_AAEL007753-RA (AAEL007753)
Expression > Probe
RNAseq reference transcript





There are currently 186 tracks turned off.
VectorBase Aedes aegypti version 83.3 (AaegL3) Supercontig supercont1.114: 1,978,029 - 1,997,740



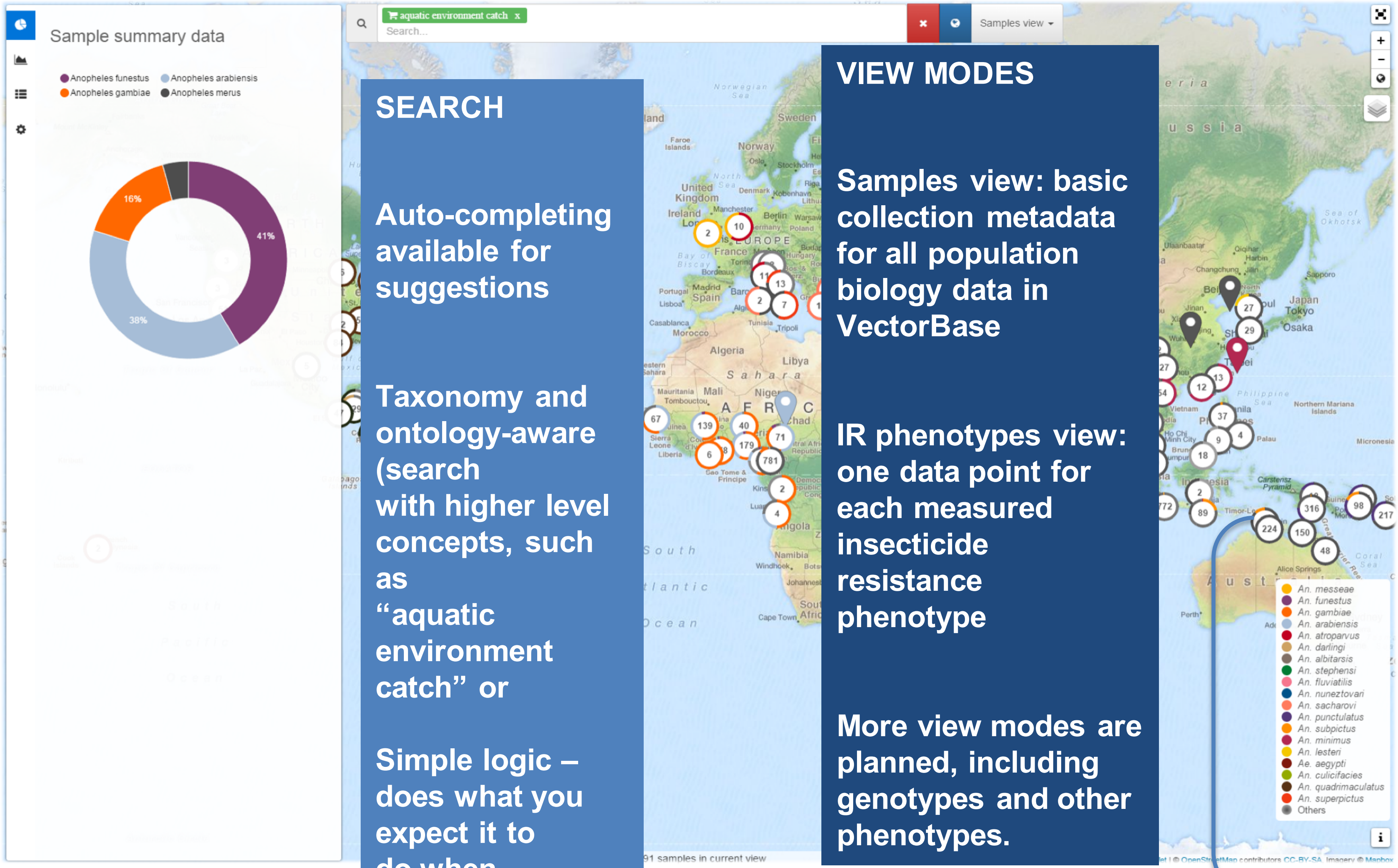
Population Biology

Welcome to VectorBase's Population Biology (PopBio) resource: a database and associated tools for visualisation, search and analysis of a wide range of population data, including genotypes, insecticide resistance and other phenotypes, and field collection metadata. You can interactively query all geo-tagged data (>99% of samples) using the map interface. Text-based search and browse is also available.

Phenotypes, genotypes and assays can be searched, browsed, and viewed in special views (see above) as we collect enough data

Large external data in resource:

- 30,000 observations from the Malaria Atlas Project
- >200 individuals with high-throughput genotyping and metadata
- 5500 insecticide resistance assays, including data from the President’s Malaria Initiative



SEARCH

Auto-completing available for suggestions

Taxonomy and ontology-aware (search with higher level concepts, such as “aquatic environment catch” or

Simple logic – does what you expect it to do when searching for several species or insecticides

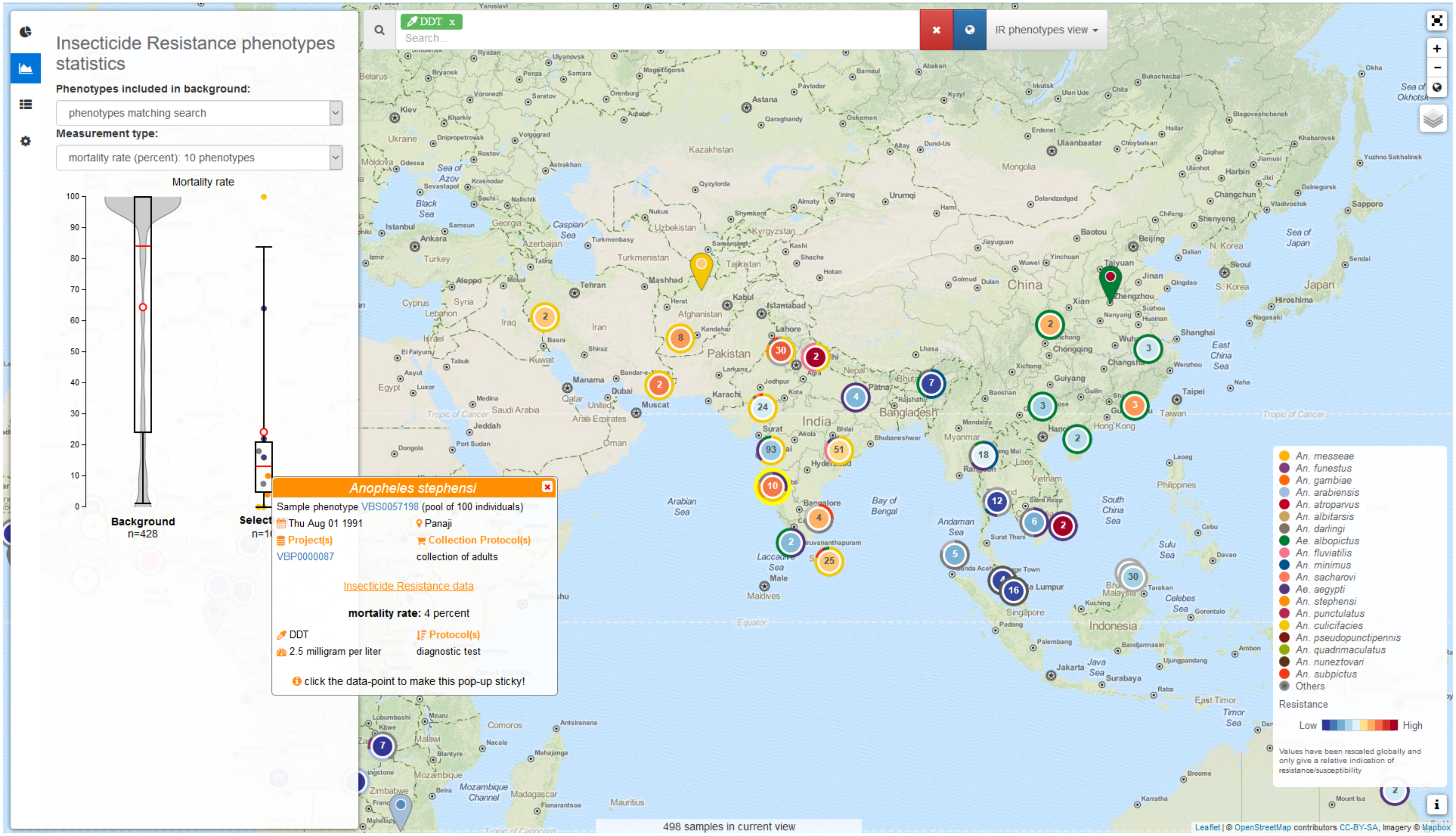
VIEW MODES

Samples view: basic collection metadata for all population biology data in VectorBase

IR phenotypes view: one data point for each measured insecticide resistance phenotype

More view modes are planned, including genotypes and other phenotypes.

Species key (color coded)



Insecticide Resistance phenotypes statistics

Phenotypes included in background: phenotypes matching search

Measurement type: mortality rate (percent): 10 phenotypes

Background
n=428

Anopheles stephensi

Sample phenotype VBS0057198 (pool of 100 individuals)

Thu Aug 01 1991 | Panaji

Project(s): VBP0000087

Collection Protocol(s): collection of adults

Insecticide Resistance data

mortality rate: 4 percent

DDT
2.5 milligram per liter

Protocol(s): diagnostic test

click the data-point to make this pop-up sticky!

Values have been rescaled globally and only give a relative indication of resistance/susceptibility

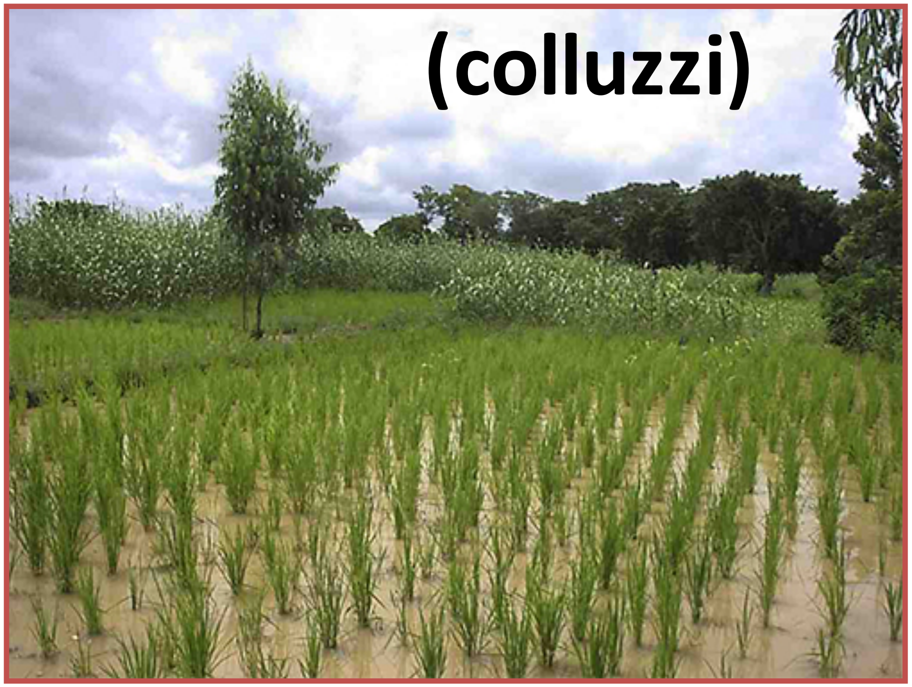
498 samples in current view

VectorBase's PopBio resource contains insecticide resistance data from a range of assay protocols and reported in a variety of measures and units, such as percent mortality, lethal concentration (e.g. LC50) and lethal time (e.g. LT95).

To aid the user in discovering geographical regions of resistance we have rescaled all comparable data, and inverting value ranges where appropriate. These rescaled values are used to color the map markers (from blue to red).

Fitness trade-offs in different habitats

M-form (colluzzi)



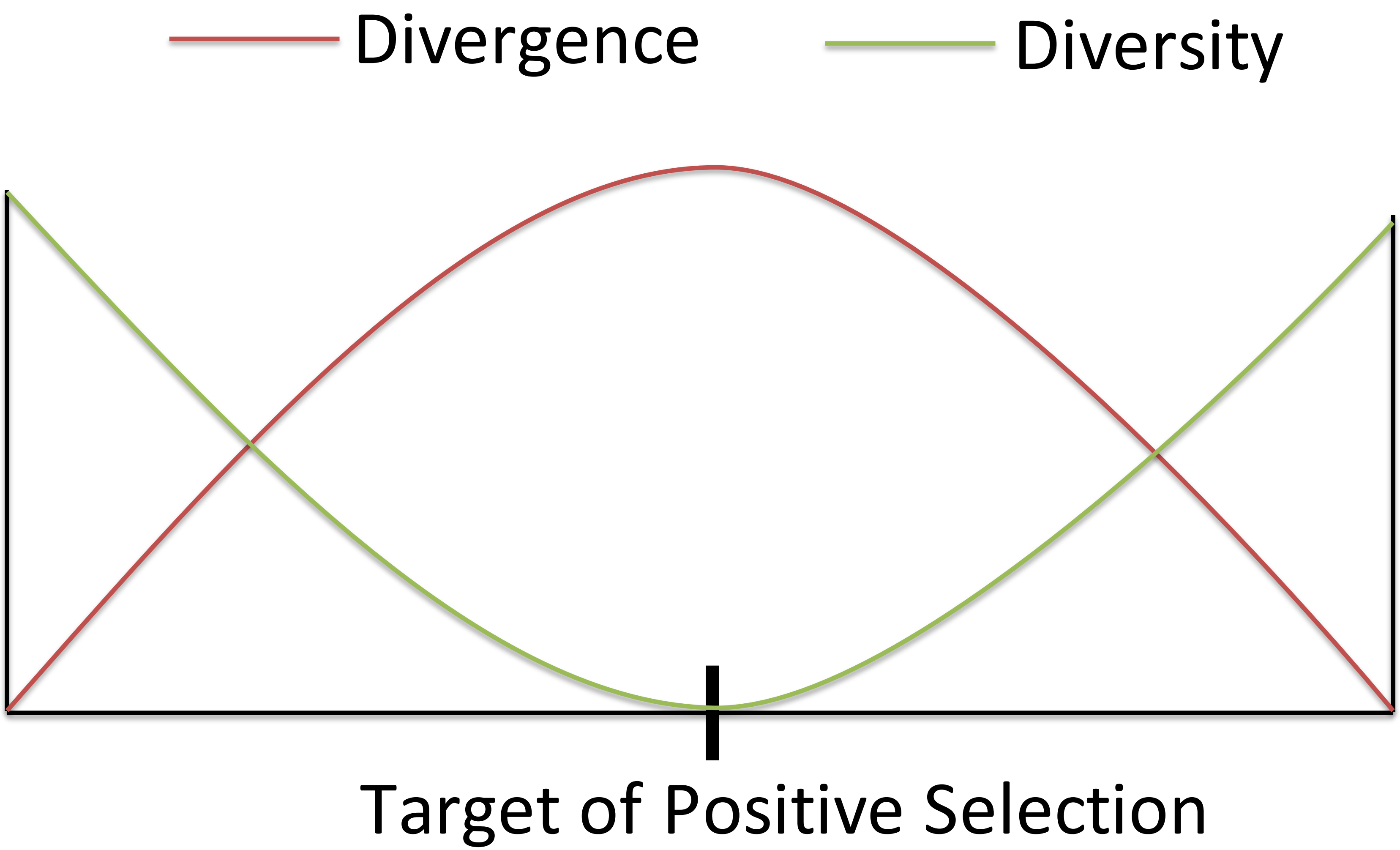
- More permanent
- Available year-round
- Allows slower development
- Predator-rich

S-form (gambiae)

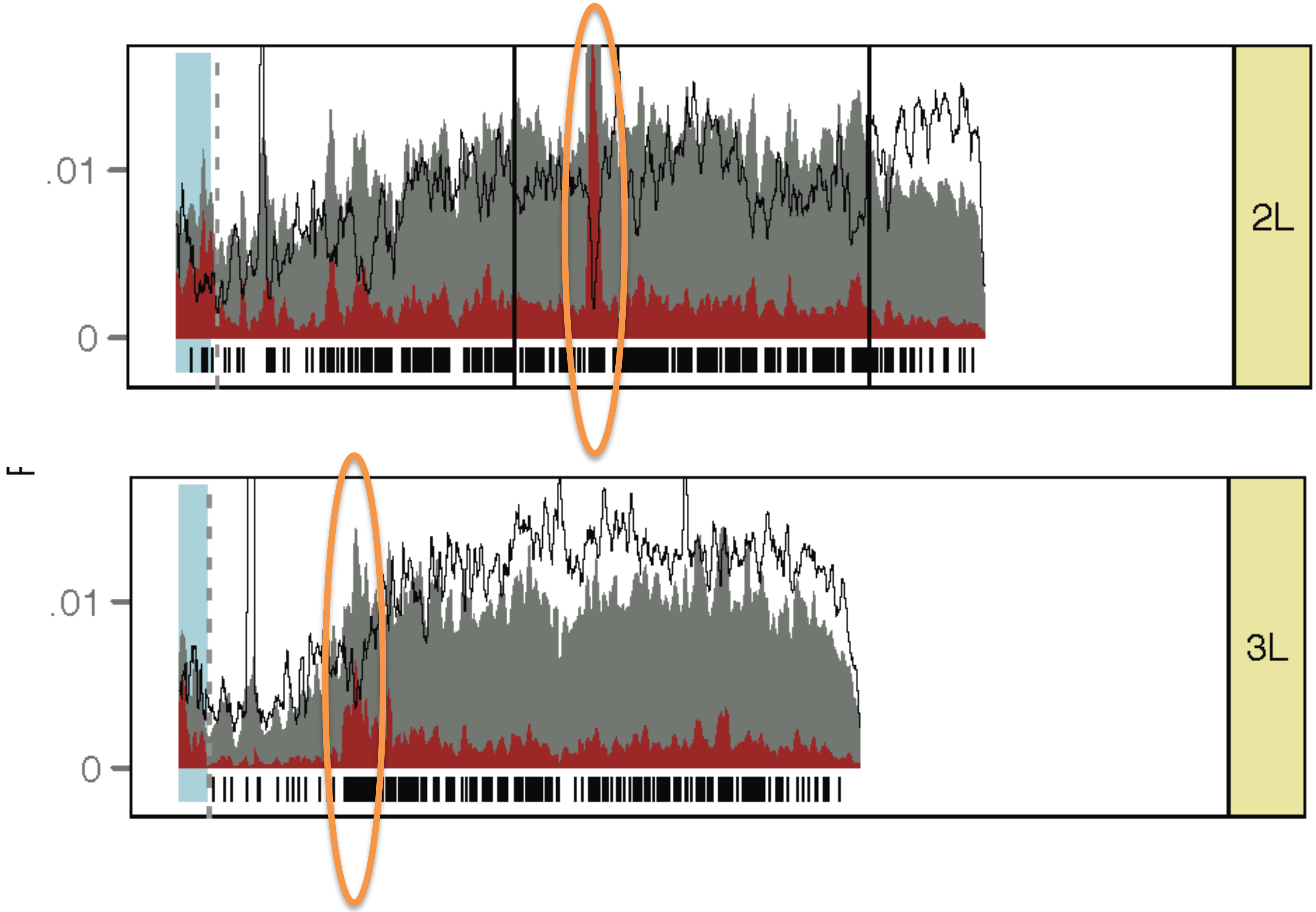
- Ephemeral
- rainy-season dependent
- Requires rapid development
- Largely predator-free

Ecological adaptation via reverse ecology

- Beneficial mutations should carry signatures

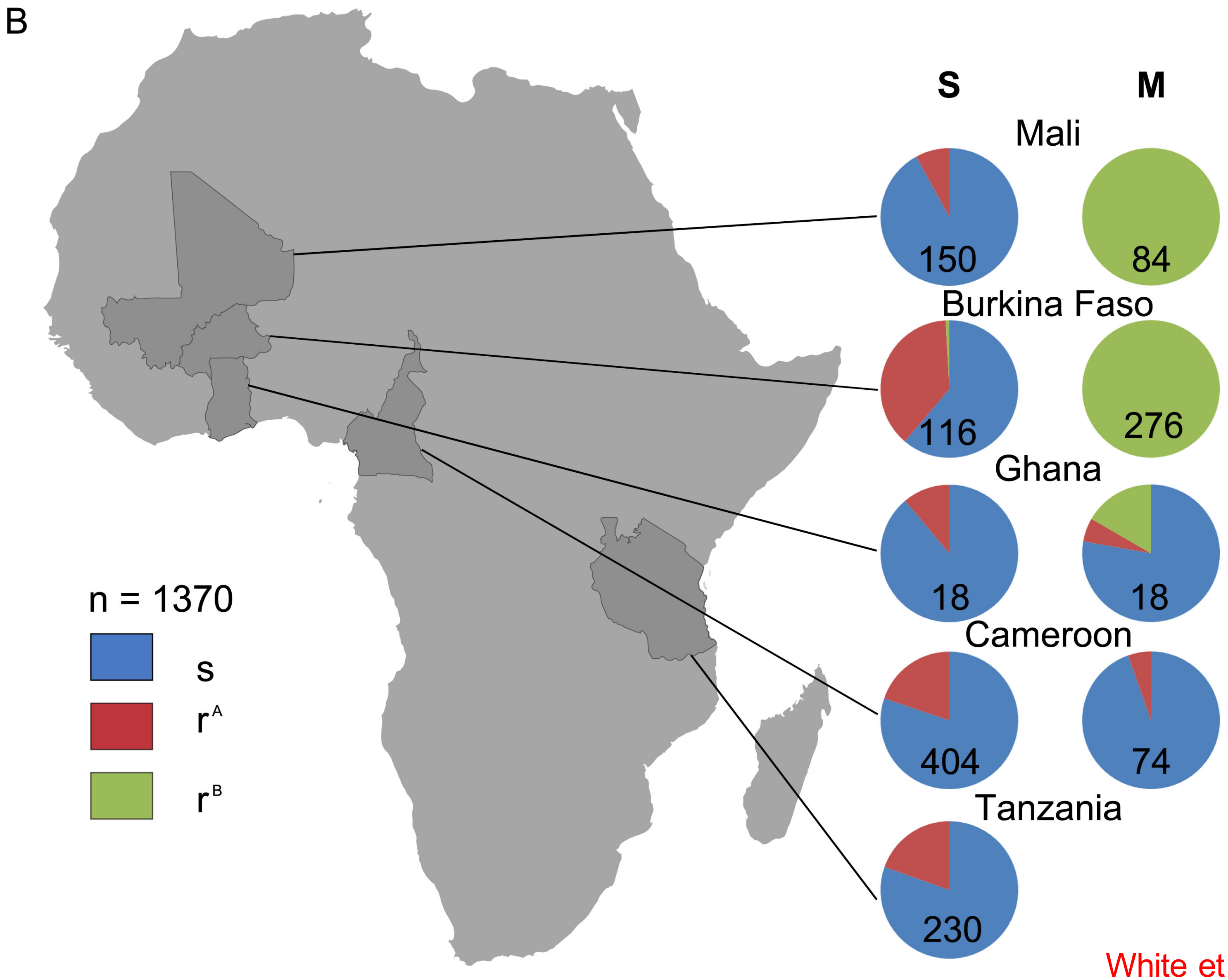


Some genomic regions display footprint of strong, recent selection



Lawniczak, Emrich et al. 2010 Science

TEP1rB is (almost) exclusive to M from West Africa



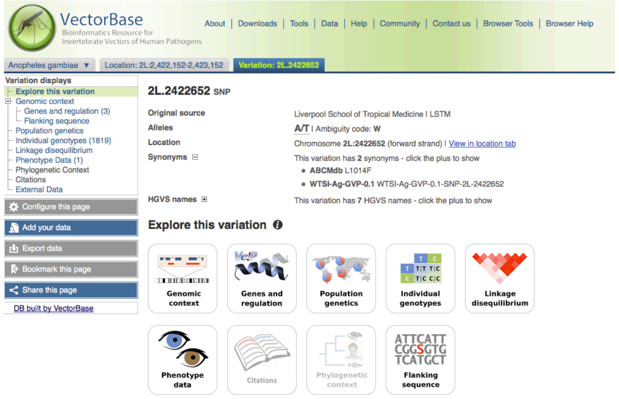
White et al. PNAS (2011)

Integrating variation and metadata across taxa

Display of variant data in genomic context

VCF

Ensembl variation database



2L.2422652 SNP

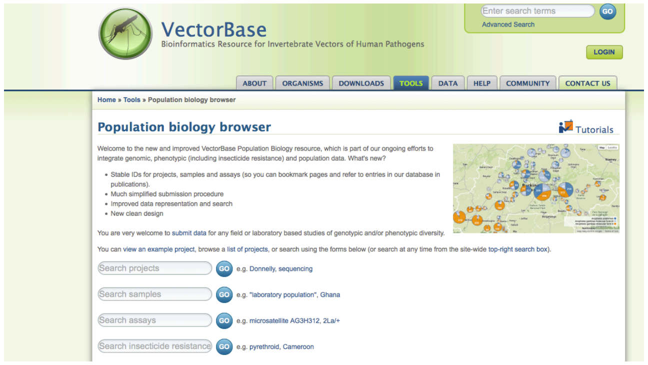
Original source	Liverpool School of Tropical Medicine I LSTM
Alleles	A/T I Ambiguity code: W
Location	Chromosome 2L:2422652 (forward strand) I View in location tab
Synonyms ⊟	This variation has 2 synonyms - click the plus to show • ABCMdb L1014F • WTSI-Ag-GVP-0.1 WTSI-Ag-GVP-0.1-SNP-2L-2422652
HGVS names ⊞	This variation has 7 HGVS names - click the plus to show

Explore this variation

Apache Solr

ISA-TAB

PopBio



Home » Tools » Population biology browser

Population biology browser

Welcome to the new and improved VectorBase Population Biology resource, which is part of our ongoing efforts to integrate genomic, phenotypic (including insecticide resistance) and population data. What's new?

- Stable IDs for projects, samples and assays (so you can bookmark pages and refer to entries in our database in publications).
- Much simplified submission procedure
- Improved data representation and search
- New clean design

You are very welcome to submit data for any field or laboratory based studies of genotypic and/or phenotypic diversity.

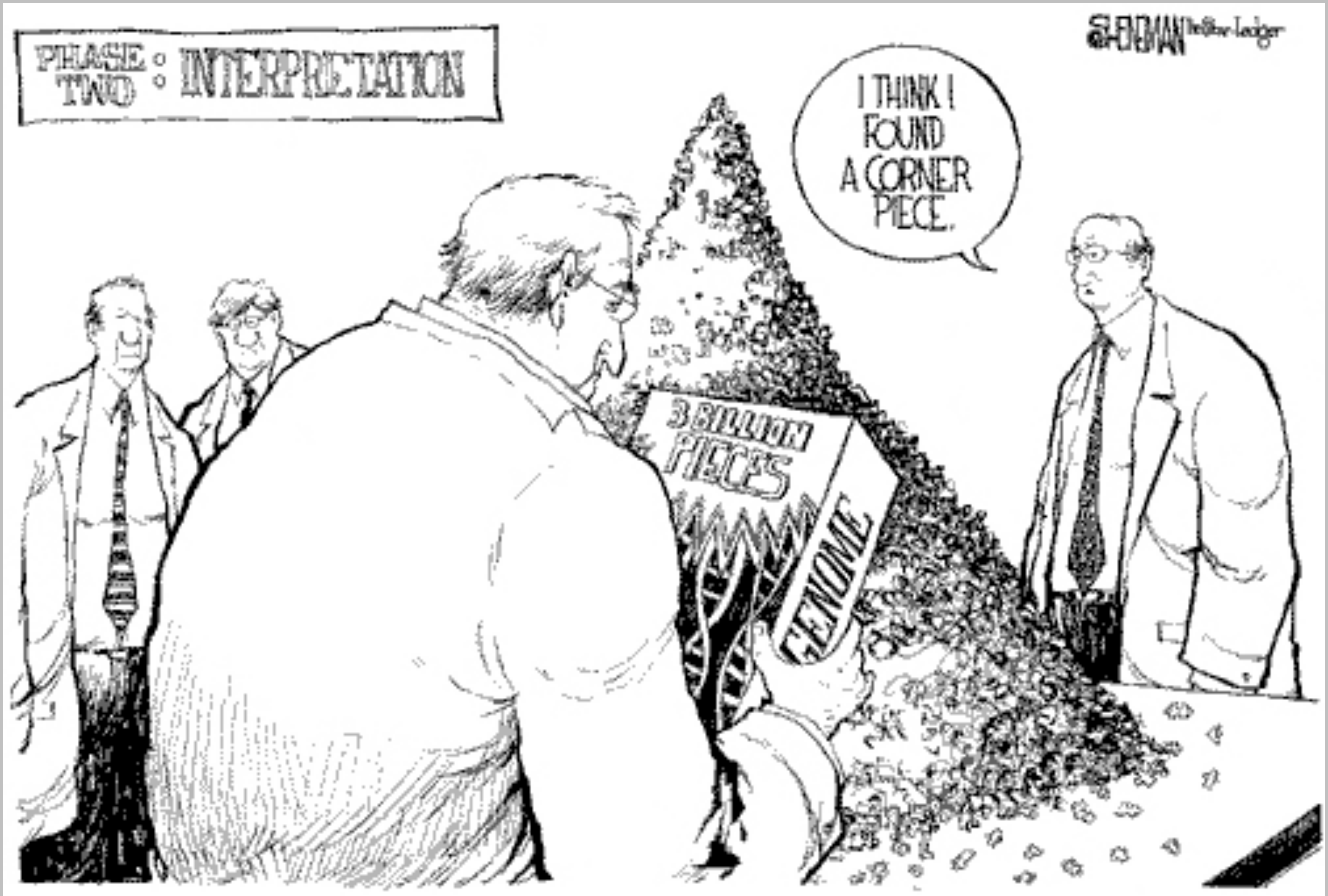
You can view an example project, browse a list of projects, or search using the forms below (or search at any time from the site-wide top-right search box).

Search projects GO e.g. Donnelly, sequencing

Search samples GO e.g. "laboratory population", Ghana

Search assays GO e.g. microsatellite AG3H312, 2La/+

Search insecticide resistance GO e.g. pyrethroid, Cameroon


PHASE TWO : INTERPRETATION
I THINK I FOUND A CORNER PIECE.
3 BILLION PIECES
GENOME
SHENEMAN The Star-Ledger

“10 simple rules” for a genome sequencing effort

- The i5K folks were most interested with these “nuggets;” note these are from my experience and therefore are my opinions

- Experience from my efforts is really the best teacher – luckily we have a great, very open community of researchers to lean on

Rule #1:
“Bioinformaticians are not ‘alchemists’”

- The quality of the manuscript is usually directly associated with the quality of
  - The data
  - The underlying questions
- Arthropod assembly still degrades with the complexity of the sample/genome
- Long reads are better but not a panacea
  - We failed to assembly *An. gambiae*’s Y and probably will continue to fail with current technology

Mostly the same methods, different results

A. freeborni

A. minimus

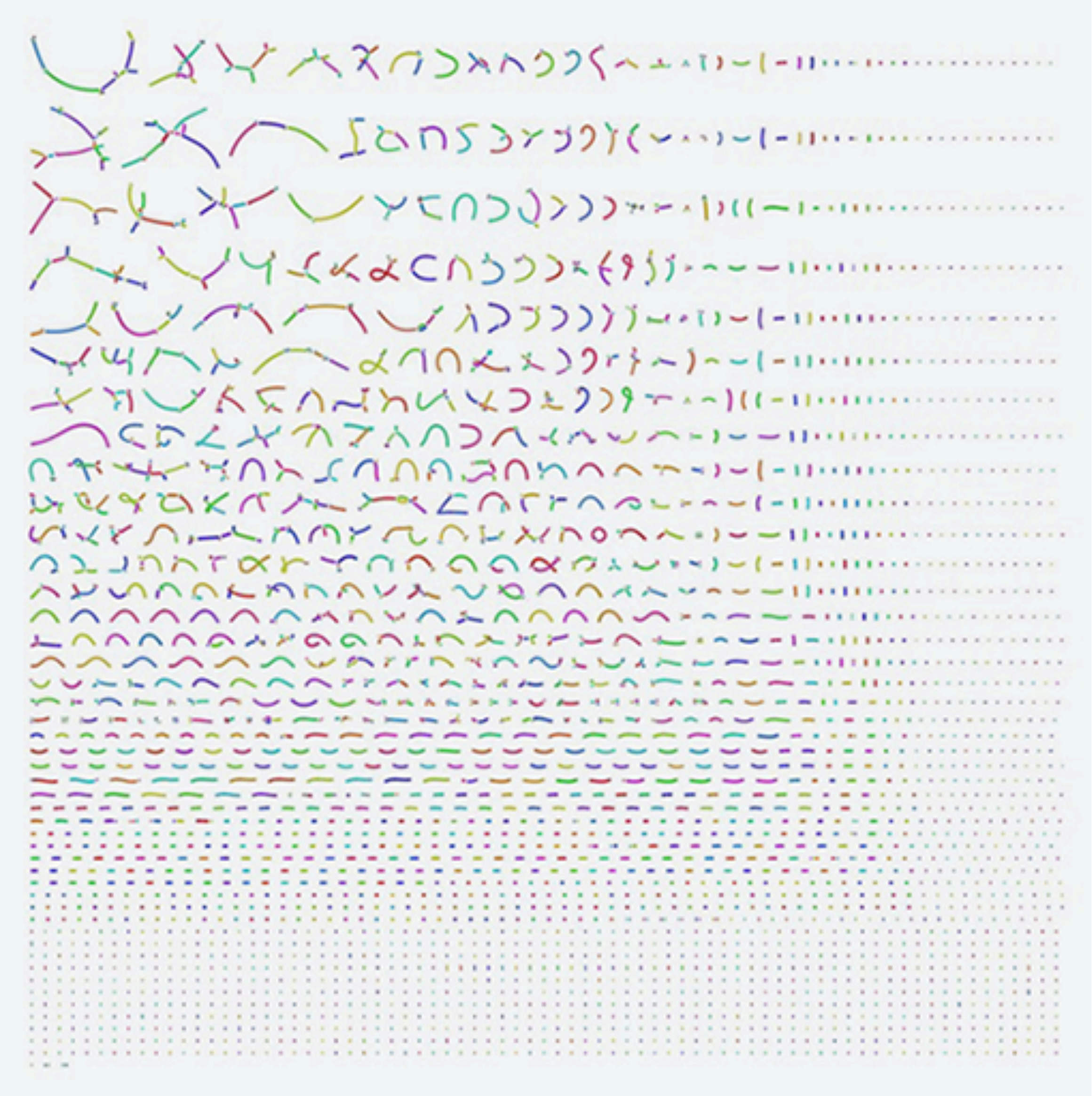
A. albimanus

Species	Genome (bp)	N50 (bp)	# Scaffolds
A. gambiae PEST	273,093,681	49,364,325	7*
A. minimus	201,793,324	10,313,149	678
A. arabiensis	246,567,867	5,604,218	1,214
A. funestus	**225,223,604**	**671,960**	**1,392**
A. stephensi	212,639,700	4,319	125,197

Ralph E. Harbach (2013). The Phylogeny and Classification of Anopheles, Anopheles mosquitoes - New insights into malaria vectors, Prof. Sylvie Manguin (Ed.), ISBN: 978-953-51-1188-7, InTech, DOI: 10.5772/54695. http://www.intechopen.com/books/anopheles-mosquitoes-new-insights-into-malaria-vectors/the-phylogeny-and-classification-of-anopheles

Rule #2:
“A good genome is a useful genome”

- No matter what the assembly method/sequencing tech, our results and others indicates quality directly correlates with heterogygosity
- In Anopheles at least, simpler libraries and DISCOVAR *de novo* can assemble the gene regions OK (Love et al., 2016)
- The corollary is hybrid approaches don’t work unless you bound variation
  - Our “protocol” from gambiae Y is to sequence full sibs



- Bandage plot of the Aedes cell line genome most recently added using PacBio (Mark Kunitomi; UCSF)

Rule #3:
"Don't be afraid to ask for help"

- Every genomics expert was at one time a novice
- We have a great community, and multiple venues exist to network/share. One coming up soon is the Arthropod Genomics Symposium to be held at Notre Dame
- Send emails as needed to your bioinformatics resource; we'd love to help!

Rule #4:
“Don’t put all of your eggs in one basket”

- The cost is usually both time and resources, but the best “flagship” manuscripts look at multiple cool biological features

- Set up working groups in your community, with a respected leader and have them write 1-2 pages with at least one figure

- You can be surprised what you find!

Rule #5:
“Use deadlines to ‘herd the cats’”

- One of the great aspects for me as a genome project veteran is how engaged people are on their favorite organism(s)

- Academics, though, are overcommitted and realized enthusiasm is variable

- Set a schedule and have the working groups present at some regular frequency; this ensures both results and a feedback loop

Rule #6:
“Teamwork can be everything”

- Some (most?) of my most successful projects involved a good friend/collaborator or two

- Examples
  - Drs. Mara Lawnziack and Alisha Holloway (M/S)
  - Dr. Matt Hahn and others (gambiae complex)
  - Drs. Scott Egan and Greg Ragland (Rhagaletis)
  - Drs. Igor Sharakov, Jake Tu, Adam Phillippy, others (gambiae Y)

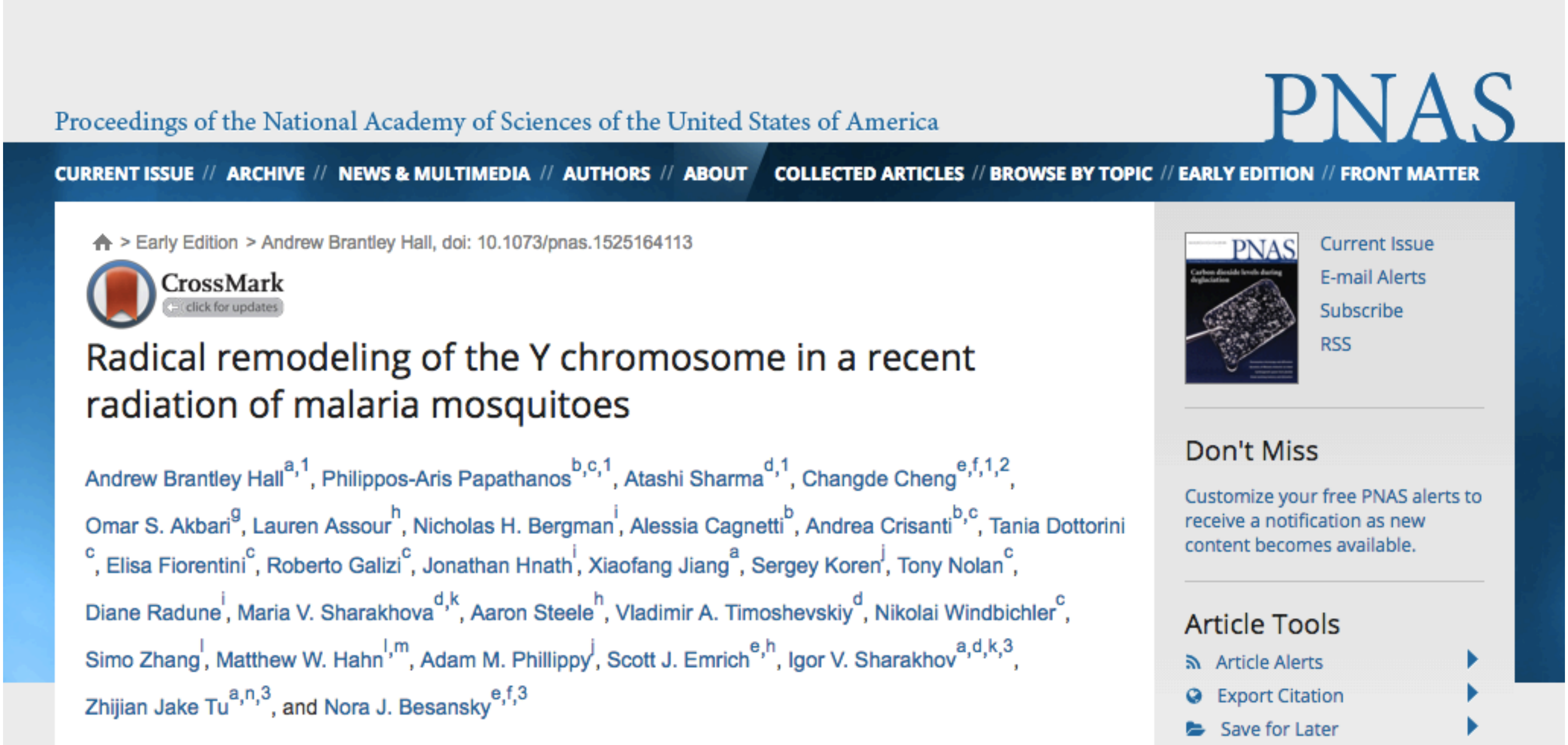
Rule #7: (KG)
“Share everything”

- Everyone knows of a horror story, but the value of open data / sharing is much higher than hoarding

- Make your genomes publically available as soon as they are frozen for the community and try and combine forces

- Archive these in national resources (e.g., GenBank)

- Your publication quality per effort expended together is much better than your group alone



Proceedings of the National Academy of Sciences of the United States of America

PNAS

CURRENT ISSUE // ARCHIVE // NEWS & MULTIMEDIA // AUTHORS // ABOUT
COLLECTED ARTICLES // BROWSE BY TOPIC // EARLY EDITION // FRONT MATTER

> Early Edition > Andrew Brantley Hall, doi: 10.1073/pnas.1525164113

CrossMark
click for updates

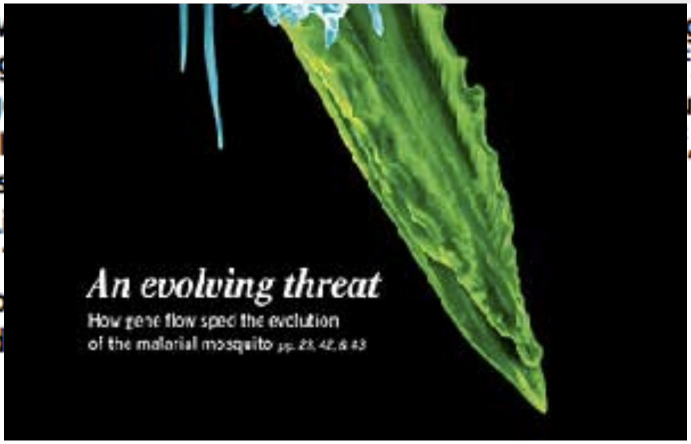
Radical remodeling of the Y chromosome in a recent radiation of malaria mosquitoes

Andrew Brantley Hall[a,1], Philippos-Aris Papathanos[b,c,1], Atashi Sharma[d,1], Changde Cheng[e,f,1,2], Omar S. Akbari[g], Lauren Assour[h], Nicholas H. Bergman[i], Alessia Cagnetti[b], Andrea Crisanti[b,c], Tania Dottorini[c], Elisa Fiorentini[c], Roberto Galizi[c], Jonathan Hnath[i], Xiaofang Jiang[a], Sergey Koren[j], Tony Nolan[c], Diane Radune[i], Maria V. Sharakhova[d,k], Aaron Steele[h], Vladimir A. Timoshevskiy[d], Nikolai Windbichler[c], Simo Zhang[l], Matthew W. Hahn[l,m], Adam M. Phillippy[j], Scott J. Emrich[e,h], Igor V. Sharakhov[a,d,k,3], Zhijian Jake Tu[a,n,3], and Nora J. Besansky[e,f,3]

PNAS
Current Issue
E-mail Alerts
Subscribe
RSS
Don't Miss
Customize your free PNAS alerts to receive a notification as new content becomes available.
Article Tools
Article Alerts
Export Citation
Save for Later

PNAS

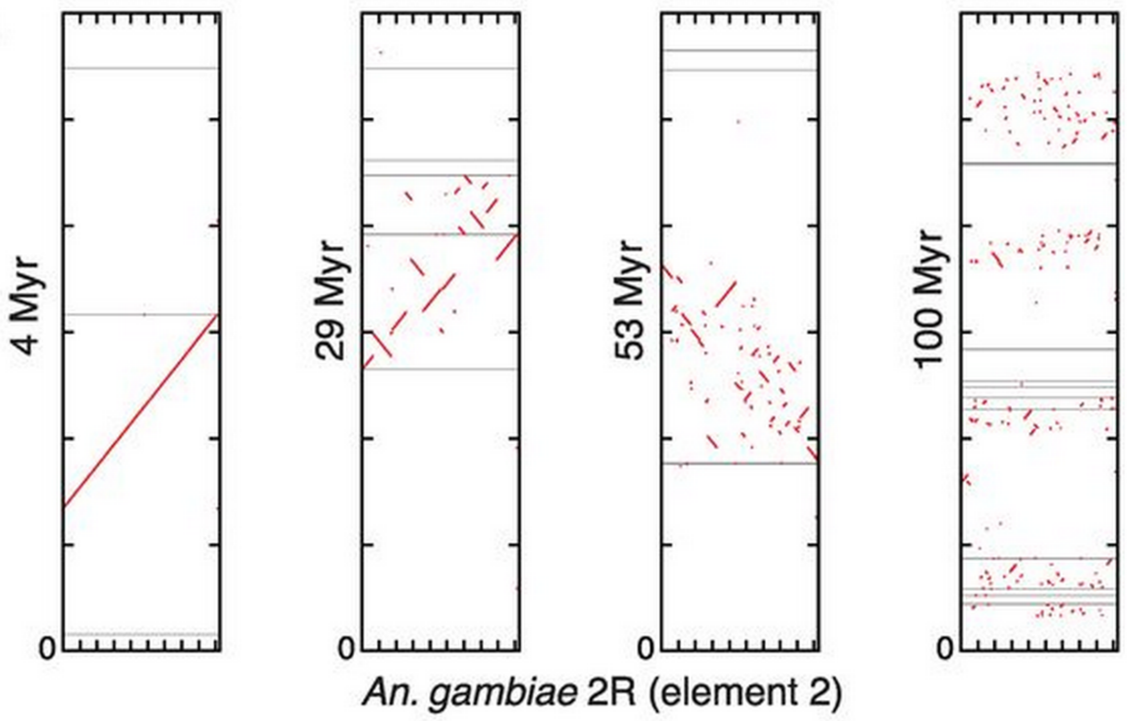
Magda D. Lugon[x], David Majerowicz[c,mm], Paula L. Marcet[kk], Marco M gy[k],
Ana C. A. Melo[a,b], Fanis Missirlis[nn], Theo Mota[oo], Fernando G. Norieg ,
Raquel L. L. Oliveira[a], Gilbert Oliveira-Silveira[c], Sheila Ons[e], Lucia Pag ual[e],
Marcio G. Pavan[g], Nicolás Pedrini[v], Alexandre A. Peixoto[b,g], Marcos ,
Francisco Prosdocimi[c], Rodrigo Ribeiro-Rodrigues[qq], Hugh M. Roberts
Didac Santesmasses[ff,gg], Renata Schama[b,g], Eloy S. Seabra-Junior[ss], L
Matheus Souza-Gomes[r], Marcos Sterkel[c], Mabel L. Taracena[c], Marta
Raul Ursic-Bedoya[d], Thiago M. Venancio[b,x], Ana Beatriz Walter-Nuno
Richard K. Wilson[h], Erwin Huebner[ww], Ellen M. Dotson[kk,2,4], and Ped



An evolving threat

How gene flow sped the evolution of the malarial mosquito pp. 23, 42, & 43

Assembly improvement using synteny (in Diptera)



- Leverage synteny to improve assemblies of phylogenetic clusters (Anopheles, Glossina)
- Test data already available (16 genomes)

Rule #8: (KG)
“Listen to the teacher(s)”

- No matter how much you share, someone will have to do the work of organizing and usually writing the paper

- This is a labor of love, but can be made easier with good junior (hungry) people leading the analysis

- There is plenty of time/opportunity for followup efforts once the first effort is done

Rule #9: (KG-ish)
“Warm food and cold beer are good for you”

- It costs resources, but the smoothest projects are ones that start over dinner and a beverage

- Try to fold in an organization meeting or two into your “big” meeting: ESA, Crete meeting, ASTMH, etc.

- An hour spent talking over beer makes much of the earlier rules much easier

Rule #10:
“Bigger is usually better”

- Using older PacBio chemistries, only local assembly continuity (contigs) is better than long-range library assisted Illumina

- The DNA demands have been large, but our best “single shot” results have been newer PacBio on size selected libraries
  - Residual haplotypes need to be removed by hand or software, esp. in inversions

- We also have had relatively good results with HiC scaffolding, which we have applied to all our proposed references (as have others)

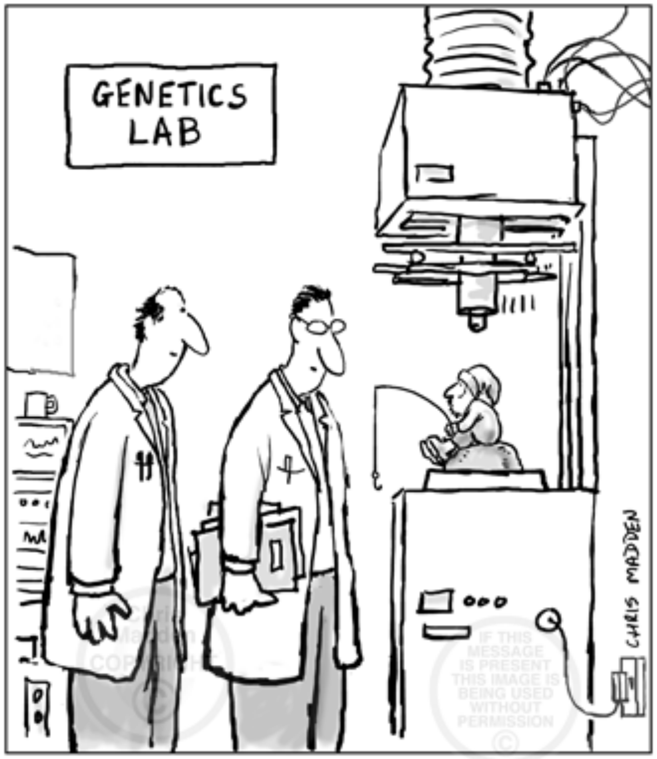
Followup and funding

Notre Dame Bioinformatics (@NDBioinformatics)

VectorBase team (EBI, ND, Imperial)

NIH/NIAID for funding

semrich@nd.edu
@ScottEmrich



THE G-NOME PROJECT

Questions?