



Every genome deserves a home:
genomic databases for the vector biologist

Dan Lawson on behalf of VectorBase
EMBL-EBI

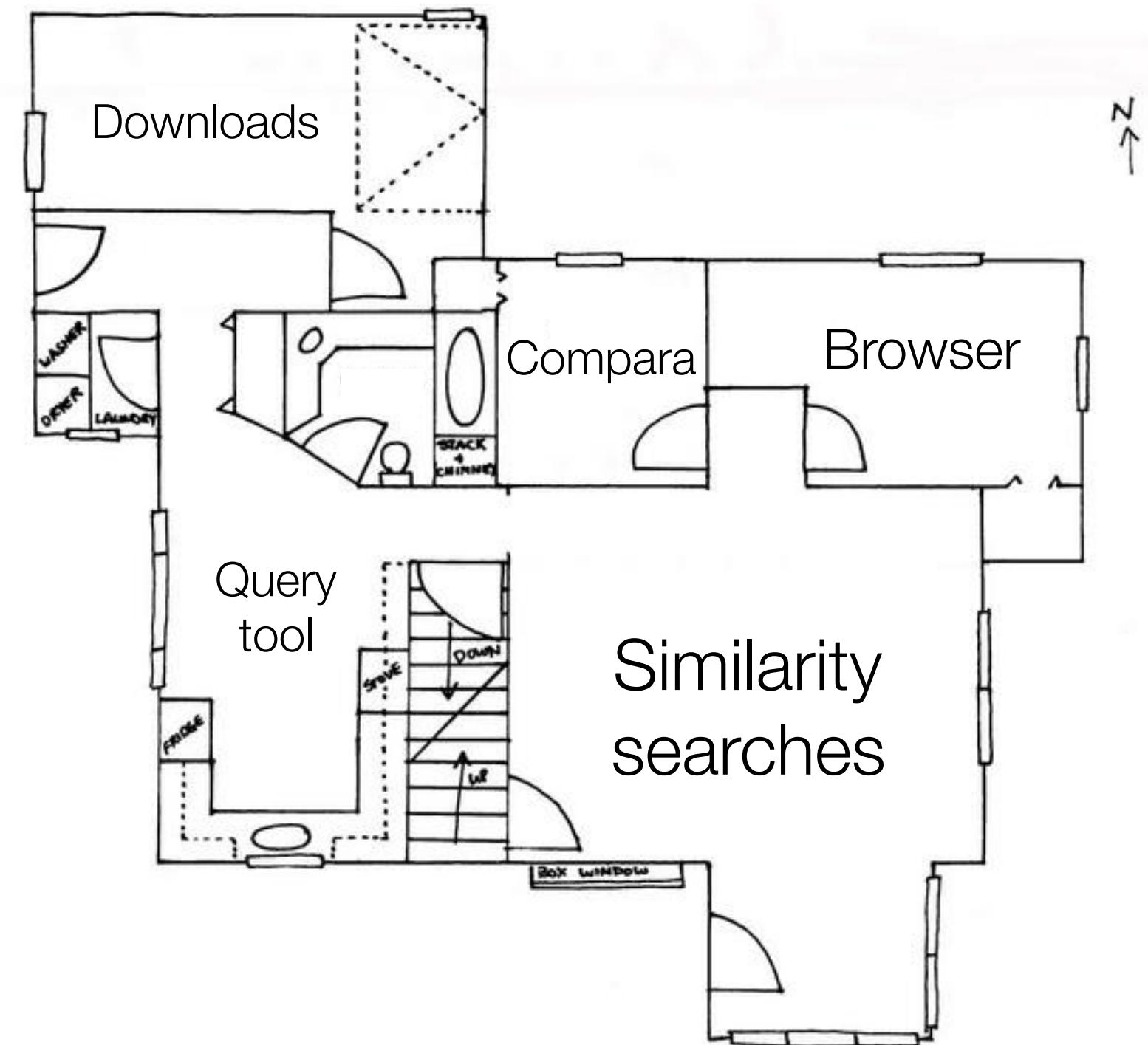


Every genome deserves a home

- Sequencing the genome of your favourite species is a beginning
- You will want to make your genome:
 - Useful to your group/community
 - Useful to other communities
- You will (hopefully) want to update/improve:
 - Assembly (new sequencing technologies, mapping strategies)
 - Gene predictions (new models, correct existing models, delete unsupported models)
 - Gene annotation (add gene names/symbols, descriptions)
 - Data richness (new high-throughput datasets, xrefs to relevant resources)

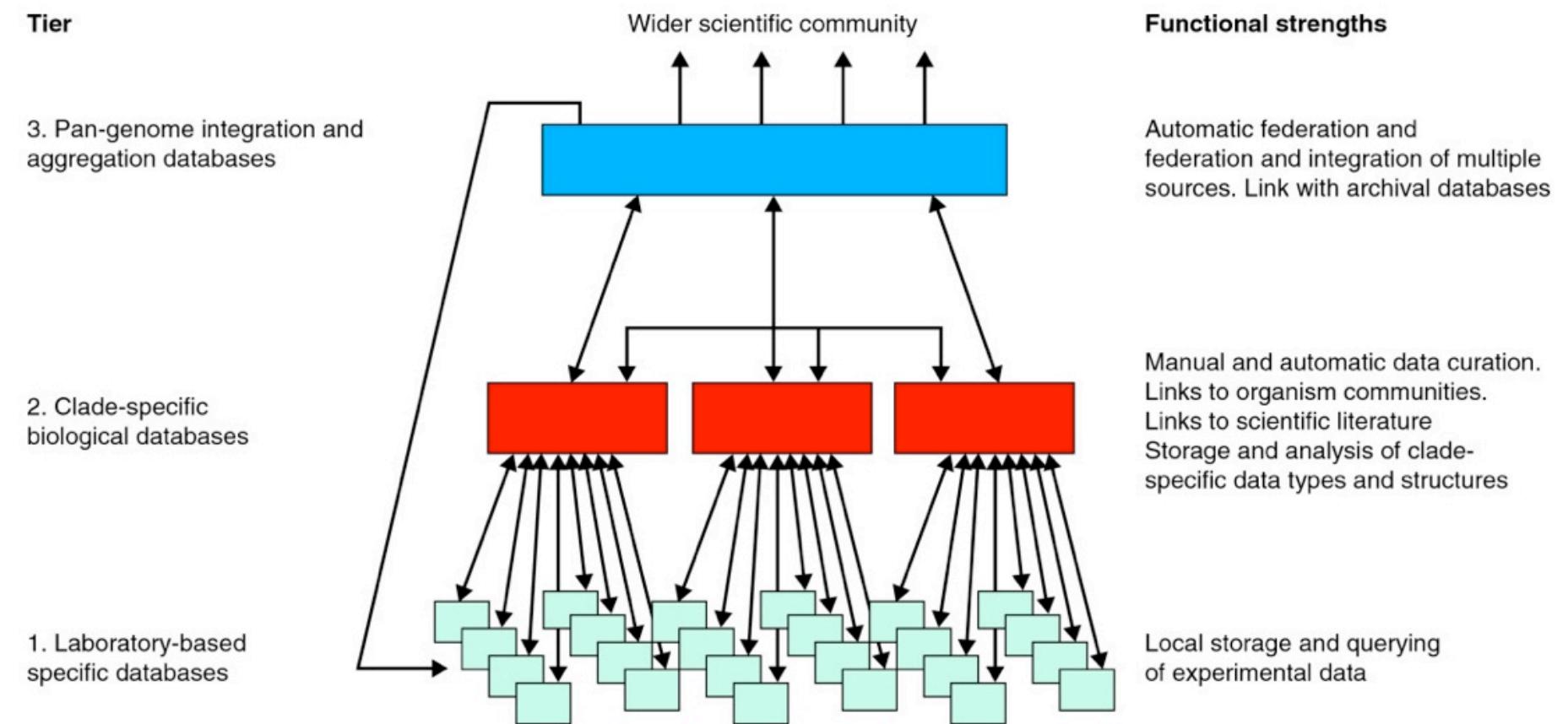
Anatomy of a home

- Genome browser
- Similarity searches
 - BLAST/BLAT
- Query tools
 - Simple keyword
 - Complex queries
- Downloads



Finding a home for every genome

- Stretching the analogy..
 - Houses
 - Apartments/Flats
 - Dormitories/Barracks



Genomic information infrastructure after the deluge

Julian Parkhill, Ewan Birney and Paul Kersey

Genome Biology 2010, 11:402

<http://genomebiology.com/2010/11/7/402>

Estate agents window

Tier 1



INSDC
UniProt
BioProject
BioSamples
GEO/ArrayExpress
Popset

Tier 2



VectorBase
Bioinformatics Resource for
Invertebrate Vectors of Human Pathogens



FlyBase
TACACAATCAGTTAGTTCCACCGACAGTCGCAGAAACCAATTGCGCGC
GTGGCAATCCGTAAGNTAGCCAATATTATTGTTAGATACTCACT
AGGTTTCAACTCAGATGTTGAGTGTAAATCAGTGAAATTC
ATTTGCGGTTTCTGAGGTTTCTGAGGTTAATGAAATGAAATGAA
ATATAATAAAACACACAGTGCAACACACGCCGGGCACTTCATAGAT
AACCTCTGCCCTGACTGGTATATGTTACTTACCATAGACATATATA

VectorBase
FlyBase
Ensembl Genomes

Tier 3



Aedes aegypti
GENE EXPRESS
UC IRVINE



Anopheles gambiae
Gene Expression Profile
UC Irvine



DRYAD



MozAtlas

(GIGA)ⁿ DB

Gene expression profiles @ UCIrvine
Aedes genome browser @ Caltech
MozAtlas, gene expression @ U. Cambridge
Popl, Population genomics @ UC Davis
Dryad/GigaDB/Supplemental data

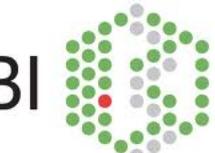
Finding a home for your data

- Visibility
 - Integration with the widest possible community
 - Cross-references back to your resource
- Longevity
 - Funding for archival databases is always going to be more secure than your database
- Accreditation
- Publication
 - Many funders and journals require submission prior to publication
 - NCBI/EBI/UCSC Browser agreement

but by far the most important thing is

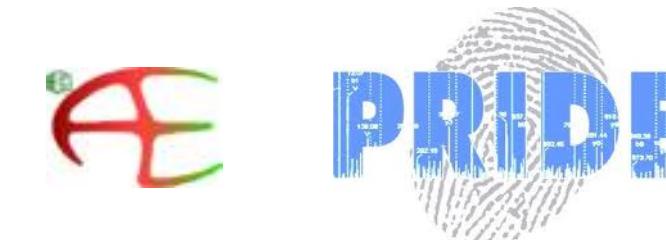
Submission to the public archival databases

EMBL-EBI



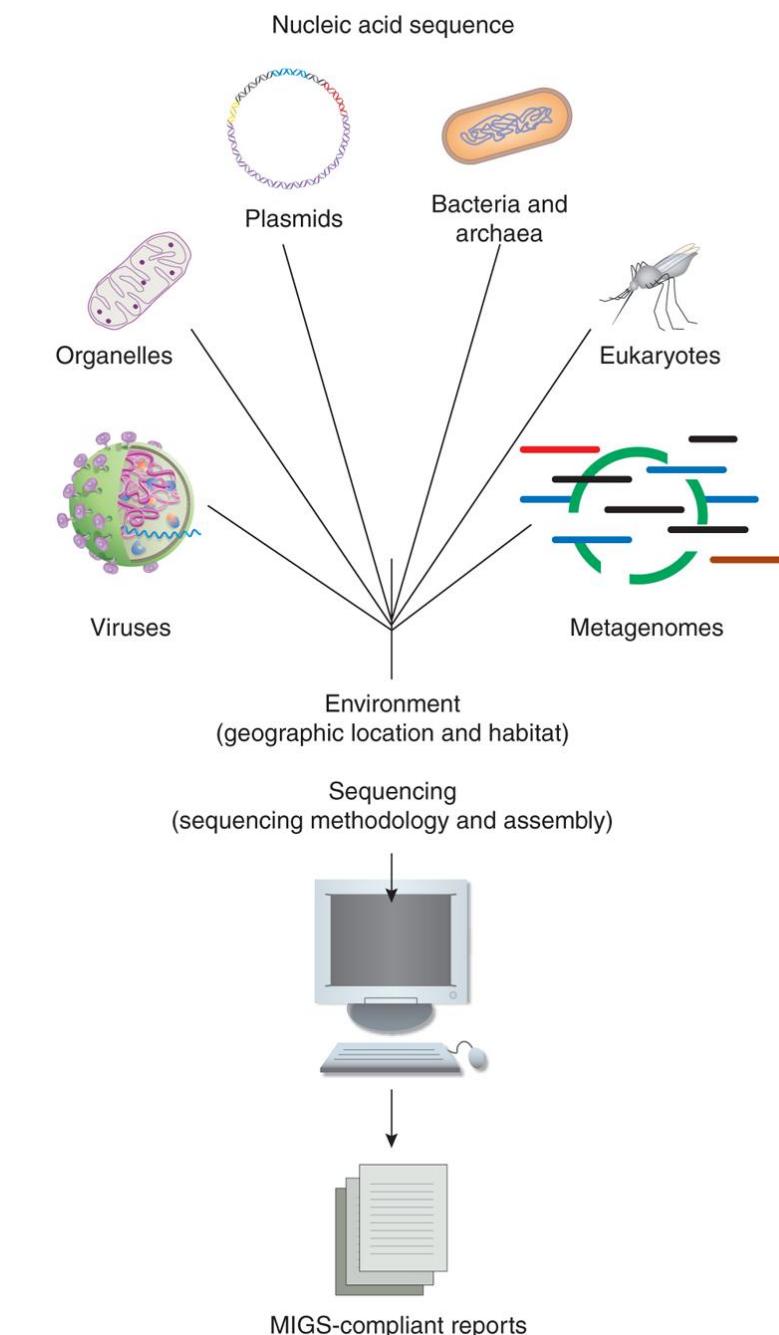
DDBJ

DNA Data Bank of Japan



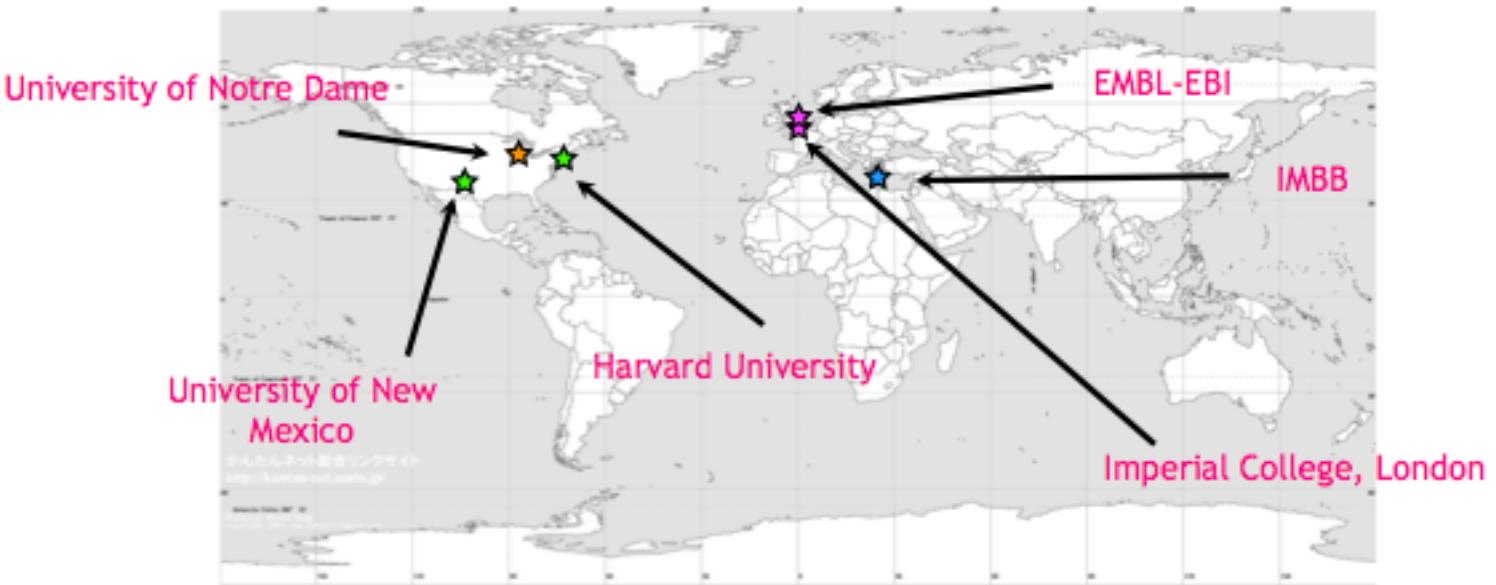
Submission makes you do a number of things

- Requirement to conform to standards
 - Some are mandatory, some advisory
 - Opportunity to capture metadata
 - Minimum information about a genome sequence (MIGS)
 - Metadata gets lost very quickly...
- Encourages good practice
- Explicit nomenclature and versioning



VectorBase: a tier 2 database

- One of 5 NIAID Bioinformatics Resource Centers
- Integrated genomic resource for arthropod vectors of human pathogens (currently 20 species)
- Collaboration of 3 European and 3 US Institutes
- VectorBase is:
 - Both service provider and content generator
 - A collator of genomic information
 - A genome annotation group (gene structure prediction)
 - A provider of tools for browsing and data mining vector genomes
 - A helpdesk for community queries
 - Responsible for data submissions to the public archival databases
 - Committed to regular release cycles (5-6 releases per year)



VectorBase (www.vectorbase.org)



- Website orientated around data rather than species
- Consolidation of legacy sections
- Faceted universal search
- Scalable handling of:
 - organism/strain
 - assembly
 - gene set
- Ensembl genome browser
- Extensive user data upload facilities
- Population Genomics
- More species
- Community Annotation Portal

Enter search terms GO

LOGOUT

VectorBase
Bioinformatics Resource for Invertebrate Vectors of Human Pathogens

ABOUT ORGANISMS DOWNLOADS TOOLS DATA HELP COMMUNITY CONTACT US ADD CONTENT

VIEW EDIT TRACK LOG DEVEL

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.

DATA

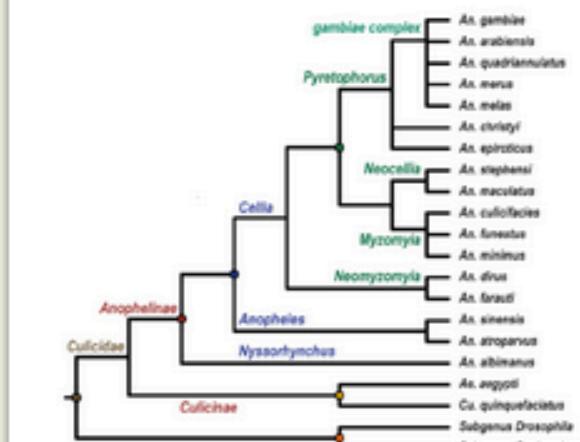
GENOMES TRANSCRIPTS & TRANSCRIPTOMES PROTEINS & PROTEOMES MITOCHONDRIAL SEQUENCES POPULATION BIOLOGY

TOOLS & RESOURCES

Resume

First pass annotation for nine Anopheline species available

VectorBase and The Anopheles Genomes Cluster announce the first pass annotation of nine Anopheline genomes. The predictions were generated using *ab initio* and similarity approaches utilising transcriptome data and taxonomically informative proteomes. Gene models were aggregated using the MAKER system. These gene sets are available for browsing, searching via BLAST and download.



An. *albimanus* An. *arabiensis*
An. *christyi* An. *dirus*
An. *epiroticus* An. *funestus*
An. *minimus* An. *quadriannulatus*

DID YOU KNOW?

- Variant Effect Predictor
Did you know you can predict the effect of variants within the vector genomes? A variant

SUBMIT YOUR DATA TO VECTORBASE

An. *stephensi* Tweets

Follow @VectorBase

DAN LAWSON Last Login: June 28, 2013

Your recent jobs

blastp 254244 254250 254319
tblastn 254367 254369 254371
blastn 255652 256063 256081

POPULAR ORGANISMS

Anopheles *gambiae* Aedes *aegypti* Culex *quinquefasciatus*

RECENT ADDITIONS

Anopheles *funestus* Phlebotomus *papatasii* Biomphalaria *glabrata*

All organisms

LATEST NEWS

June 28, 2013
VectorBase Release VB-2013-06
May 3, 2013
Rhodnius prolixus gene set
RproC1.1 released

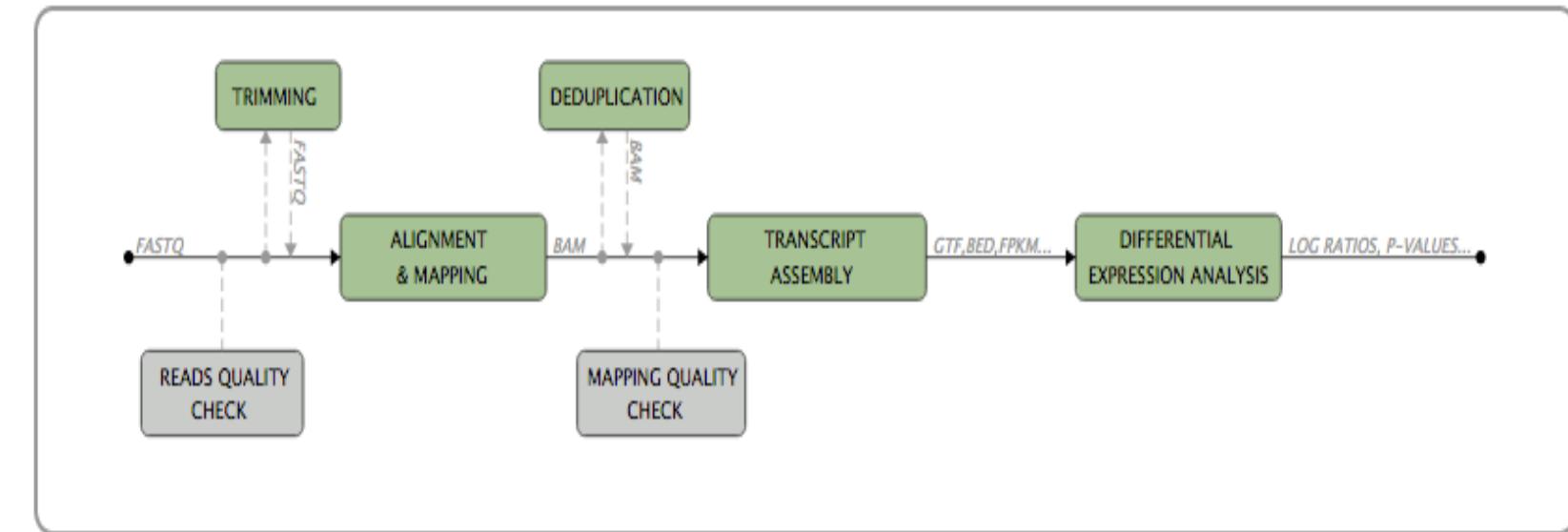
More news

VectorBase (www.vectorbase.org)



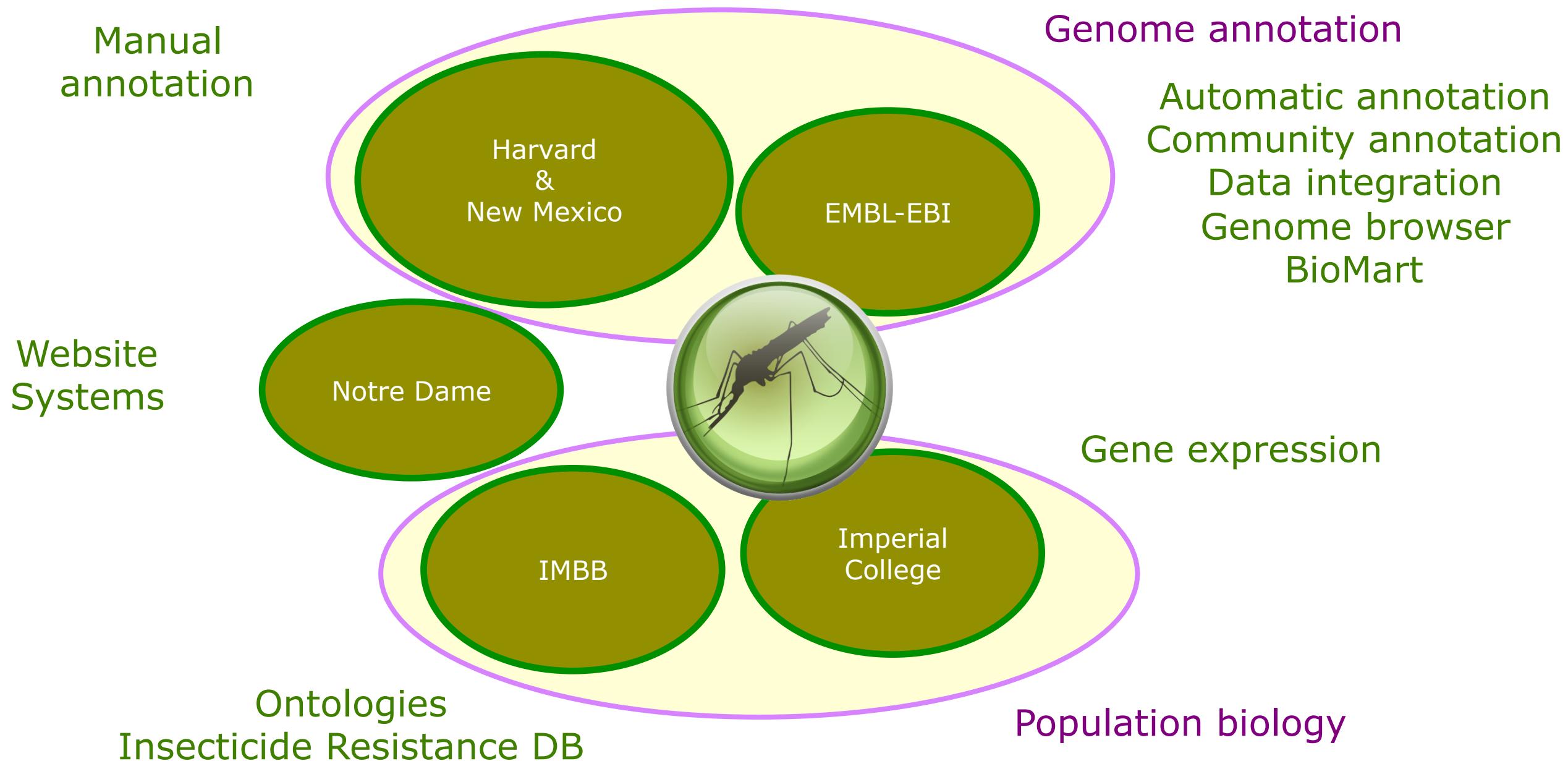
- Website orientated around data rather than species
- Consolidation of legacy sections
- Faceted universal search
- Scalable handling of:
 - organism/strain
 - assembly
 - gene set
- Ensembl genome browser
- Extensive user data upload facilities
- Population Genomics
- More species
- Community Annotation Portal

RNA-Rocket



<http://rnaseq.pathogenportal.org/>

VectorBase project roles



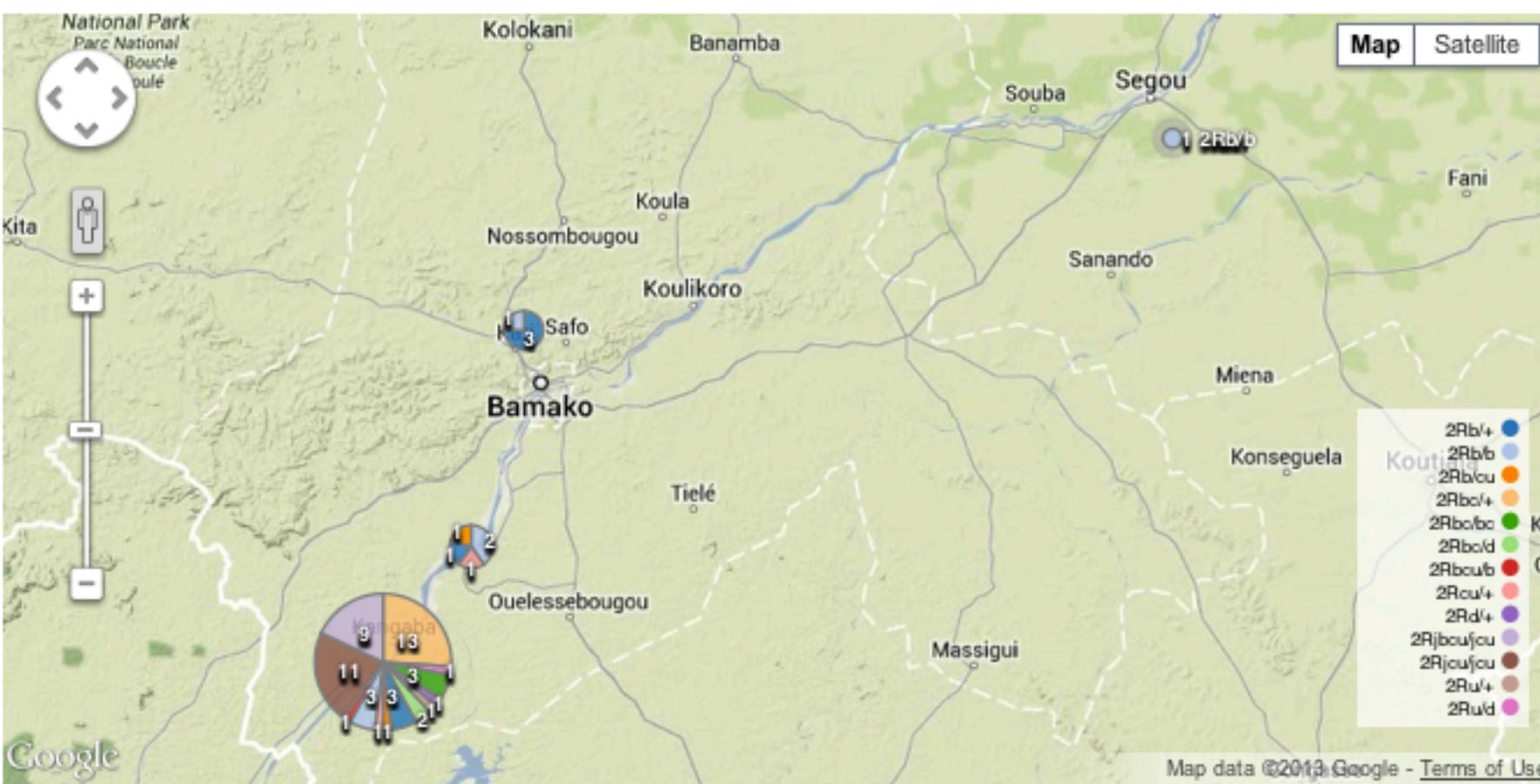
Population biology resource

- Built on ontologies
- Consolidate the following:
 - Surveillance studies (e.g. species, molecular forms)
 - Population structure studies (e.g. karyotype, microsatellite data)
 - Insecticide resistance studies
 - Metadata for high-throughput re-sequencing and SNP calling
- Internal links to genome browser
- External links to BioSamples, SRA

The screenshot shows the VectorBase Population biology browser. At the top, there's a navigation bar with links for ABOUT, ORGANISMS, DOWNLOADS, TOOLS (which is highlighted in blue), DATA, HELP, COMMUNITY, CONTACT US, and ADD CONTENT. A search bar at the top right contains the placeholder "Enter search terms" with a "GO" button. Below the navigation bar, a green banner reads "VectorBase Bioinformatics Resource for Invertebrate Vectors of Human Pathogens". The main content area has a header "Population biology browser". It features a map of a river system with a pie chart overlay showing data for locations like Banjoumene and Kengaha. To the right of the map is a legend with color-coded entries: 21fb1x (blue), 21fb2b (orange), 21fb2c (green), 21fb2d (red), 21fb2e (purple), 21fb2f (yellow), 21fb2g (pink), 21fb2h (brown), 21fb2i (light blue), and 21fb2j (grey). Below the map, there are four search boxes: "Search projects" (e.g. Donnelly, sequencing), "Search samples" (e.g. "laboratory population", Ghana), "Search assays" (e.g. microsatellite AG3H312, 2La/+), and "Search insecticide resistance" (e.g. pyrethroid, Cameroon). At the bottom, there are logos for the University of Notre Dame, EMBL-EBI, IMBB, Harvard University, Imperial College London, and The University of New Mexico. A footer note states: "This project has been funded in whole or in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services."

Example project: Species and karyotype data

- Project metadata
 - Graphical summaries of data (maps, charts)
 - Sample information with links to Biosamples (where available)



VectorBase

Bioinformatics Resource for Invertebrate Vectors of Human Pathogens

Enter search terms

GO

LOGIN

ABOUT ORGANISMS DOWNLOADS TOOLS DATA HELP COMMUNITY CONTACT US

Home » Tools » Population biology browser » Projects

Project

VectorBase stable ID	VBP0000003
Name	Neafsey et al., 2010 Anopheles gambiae M, S and Bamako populations - AgSNP01 genotypes and karyotypes for chromosome arms 2L and 2R
Description	Female mosquitoes were collected from Mali via spray catch in 2004. These were identified, using a combination of cytological karyotyping and molecular assays, as one of the three known sympatric A. gambiae populations in Mali (M, S, and Bamako). Genotyping at ~60k quality-controlled loci was performed using the Affymetrix SNP chip (AgSNP01).
Contact(s)	Nora Besansky (University of Notre Dame), Seth Redmond (Imperial College London)
Submission date	2011
Public release date	2011
Last modified date	2013-06-26
study design	observational design
study design	strain or line design
study design	genotype design

Publications

SNP genotyping defines complex gene-flow boundaries among African malaria vector mosquitoes

Neafsey DE, Lawniczak MK, Park DJ, Redmond SN, Coulibaly MB, Traoré SF, Sagnon N, Costantini C, Johnson C, Wiegand RC, Collins FH, Lander ES, Wirth DF, Kafatos FC, Besansky NJ, Christophides GK, Muskavitch MA.
published PubMed

Graphical summaries

Map data ©2013 Google - Terms of Use

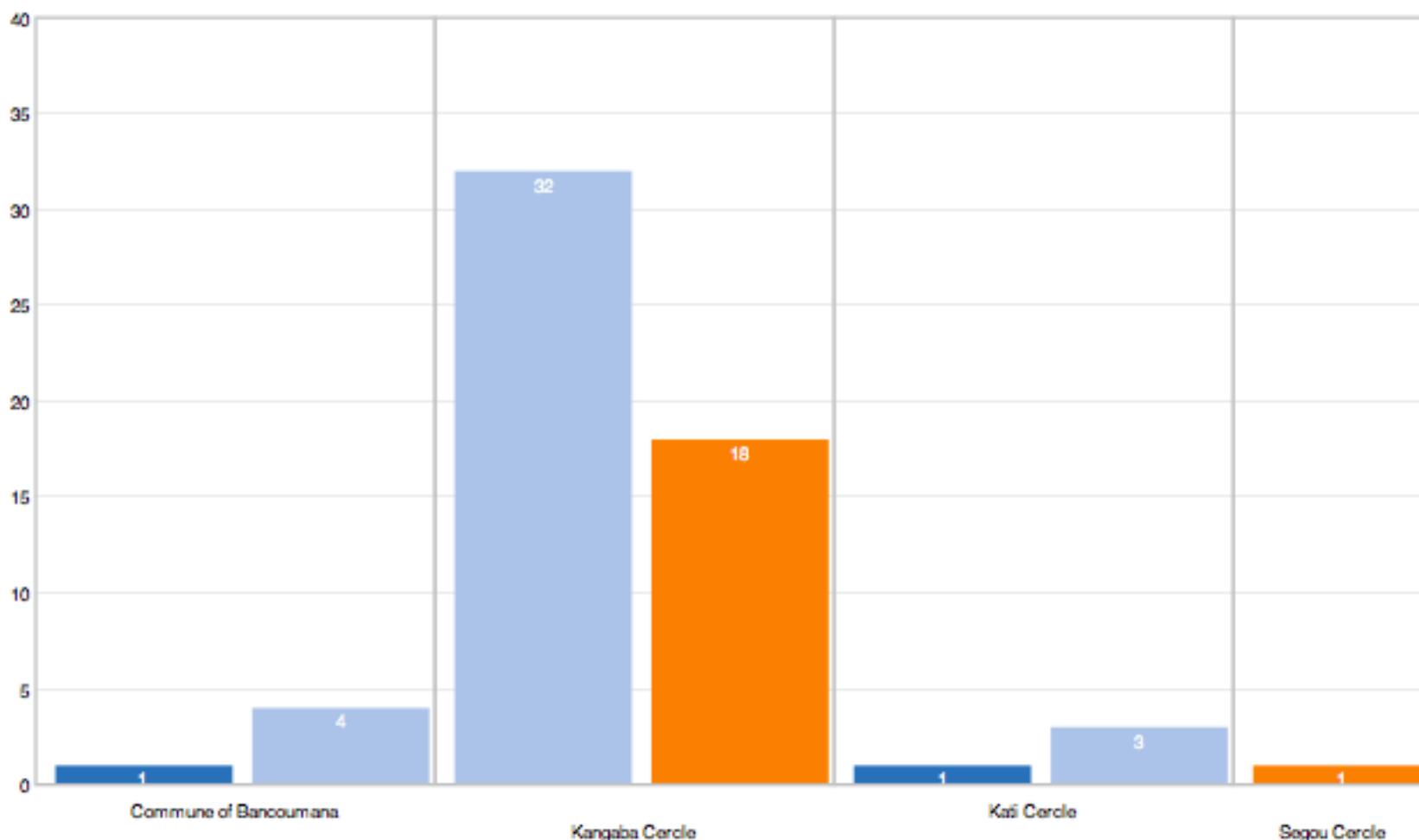
Samples

FIRST ←PREV 1 TO 20 OF 60 NEXT→ LAST

Sample	Species	Properties
SM-NRUE (VBS0000073)	Anopheles gambiae molecular form M	material type: individual sex: female developmental stage: adult comment [BioSamples accession]: SAMEA568610
SM-NRUF (VBS0000074)	Anopheles gambiae molecular form M	material type: individual sex: female developmental stage: adult comment [BioSamples accession]: SAMEA568643
SM-NRUG (VBS0000075)	Anopheles gambiae molecular form M	material type: individual sex: female developmental stage: adult comment [BioSamples accession]: SAMEA568629

Example project: Species and karyotype data

- Project metadata
- Graphical summaries of data (maps, charts)
- Sample information with links to BioSamples (where available)



The screenshot displays the VectorBase Population biology browser interface for a specific project. The top navigation bar includes links for ABOUT, ORGANISMS, DOWNLOADS, TOOLS (highlighted in blue), DATA, HELP, COMMUNITY, and CONTACT US. The search bar at the top right contains the placeholder "Enter search terms".

Project

VectorBase stable ID: VBP0000003
Name: Neafsey et al., 2010 Anopheles gambiae M, S and Bamako populations - AgSNP01 genotypes and karyotypes for chromosomes arms 2L and 2R
Description: Female mosquitoes were collected from Mali via spray catch in 2004. These were identified, using a combination of cytological karyotyping and molecular assays, as one of the three known sympatric A. gambiae populations in Mali (M, S, and Bamako). Genotyping at ~60k quality-controlled loci was performed using the Affymetrix SNP chip (AgSNP01).
Contact(s): Nora Besansky (University of Notre Dame), Seth Redmond (Imperial College London)
Submission date: 2011
Public release date: 2011
Last modified date: 2013-06-26
study design: observational design
study design: strain or line design
study design: genotype design

Publications

SNP genotyping defines complex gene-flow boundaries among African malaria vector mosquitoes
Neafsey DE, Lawtonchuk MK, Park DJ, Redmond SN, Coulibaly MB, Traoré SF, Sagnon N, Costantini C, Johnson C, Wiegand RC, Collins FH, Lander ES, Wirth DF, Kafatos FC, Besansky NJ, Christophides GK, Muskavitch MA.
published PubMed

Graphical summaries

Species distribution, 2L inversion karyotype, 2R inversion karyotype, 2L karyotype by collection site

Samples

FIRST ←PREV 1 TO 20 OF 60 NEXT→ LAST

Sample	Species	Properties
SM-NRUE (VBS0000073)	Anopheles gambiae molecular form M	material type: individual sex: female developmental stage: adult comment [BioSamples accession]: SAMEA568610
SM-NRUF (VBS0000074)	Anopheles gambiae molecular form M	material type: individual sex: female developmental stage: adult comment [BioSamples accession]: SAMEA568643
SM-NRUG (VBS0000075)	Anopheles gambiae molecular form M	material type: individual sex: female developmental stage: adult comment [BioSamples accession]: SAMEA568629

Example sample: NGS re-sequencing & SNP calling

- Sample metadata
- Annotated using ontologies (both VectorBase generated, e.g. MIRO and external e.g. EFO, PATO)
- Links to parental project
- Links to data (BioSamples/SRA)

Enter search terms

 **VectorBase**
Bioinformatics Resource for Invertebrate Vectors of Human Pathogens

[ABOUT](#) [ORGANISMS](#) [DOWNLOADS](#) **TOOLS** [DATA](#) [HELP](#) [COMMUNITY](#) [CONTACT US](#)

Home » Tools » Population biology browser » Samples

Sample

VectorBase stable ID	VBS0000059
Name	AK0035-C
Description	individual specimen from Kisumu colony (MR4)
Species	<i>Anopheles gambiae</i> (derived, unambiguous)
Material type	individual
sex	female
developmental stage	adult
comment [colony]	Kisumu
comment [Ensembl variation ID]	1926
comment [BioSamples accession]	SAMEA838737

Assays

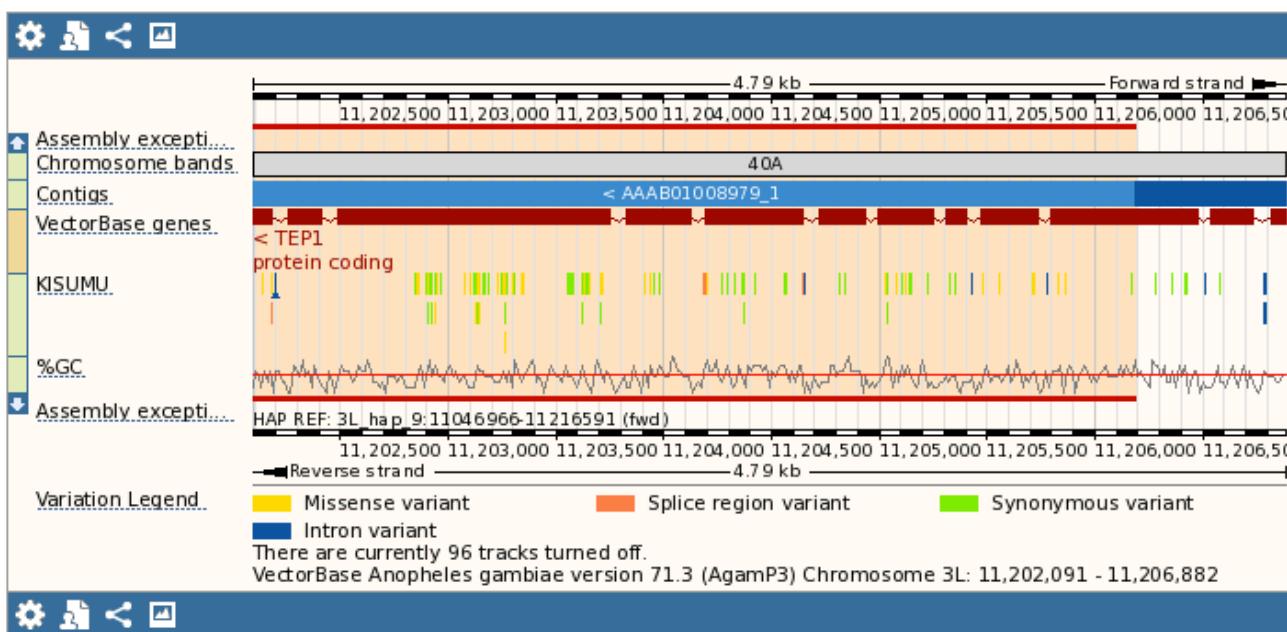
Assay	Summary	Project(s)
Genotype assay AK0035-C.seq1 VBA0000074	variants shown in genome browser (genotyping by high throughput sequencing)	 VBP0000002
Sample manipulation AK0035-C derived from VBS0000026 VBA0000060	basic sample manipulation	VBP0000002

This sample belongs to project(s)

Sequencing studies of MR4 derived insect colonies (VBP0000002)
These samples were sequenced by the Wellcome Trust Sanger Institute, as part of the Malaria Programme's *Anopheles gambiae* Genome Variation Project (linked as http://www.sanger.ac.uk/research/projects/malariaprogramme-kwiatkowski/#_research). The samples were provided by the Malaria Research and Reference Reagent Resource Center (www.mr4.org)
Submitted: 2013-04-17
Contact(s): Dominic Kwiatkowski (Wellcome Trust Sanger Institute), Martin Donnelly (Wellcome Trust Sanger Institute), James Stalker (Wellcome Trust Sanger Institute)

Example sample: NGS re-sequencing & SNP calling

- Sample metadata
- Annotated using ontologies (both VectorBase generated, e.g. MIRO and external e.g. EFO, PATO)
- Links to parental project
- Links to data (BioSamples/SRA)



Sample

VectorBase stable ID	VBS0000059
Name	AK0035-C
Description	individual specimen from Kisumu colony (MR4)
Species	<i>Anopheles gambiae</i> (derived, unambiguous)
Material type	individual
sex	female
developmental stage	adult
comment [colony]	Kisumu
comment [Ensembl variation ID]	1926
comment [BioSamples accession]	SAMEA838737

Assays

Assay	Summary	Project(s)
Genotype assay AK0035-C.seq1 VBA0000074	variants shown in genome browser (genotyping by high throughput sequencing)	VBP000002
Sample manipulation AK0035-C derived from VBS0000026 VBA0000080	basic sample manipulation	VBP000002

This sample belongs to project(s)

Sequencing studies of MR4 derived insect colonies (VBP000002)
These samples were sequenced by the Wellcome Trust Sanger Institute, as part of the Malaria Programme's *Anopheles gambiae* Genome Variation Project (linked as http://www.sanger.ac.uk/research/projects/malaria-programme-kwiatkowski/#_research). The samples were provided by the Malaria Research and Reference Reagent Resource Center (www.mr4.org)

Submitted: 2013-04-17
Contact(s): Dominic Kwiatkowski (Wellcome Trust Sanger Institute), Martin Donnelly (Wellcome Trust Sanger Institute), James Stalker (Wellcome Trust Sanger Institute)

Population genetics visualisation

 **VectorBase**
Bioinformatics Resource for
Invertebrate Vectors of Human Pathogens

Anopheles gambiae ▾ Location: 2L:2,422,152-2,423,152 Variation: 2L:2422652

Variation displays

- Explore this variation
- Genomic context
 - Genes and regulation (3)
 - Flanking sequence
 - Population genetics
 - Individual genotypes (1819)
 - Linkage disequilibrium
 - Phenotype Data (1)
 - Phylogenetic Context (1)
 - External Data
- Configure this page
- Add your data
- Export data
- Bookmark this page
- Share this page
- DB built by VectorBase

2L.2422652 SNP

Original source: Liverpool School of Tropical Medicine | LSTM

Alleles: Reference/Alternative: A/T | Ambiguity code: W

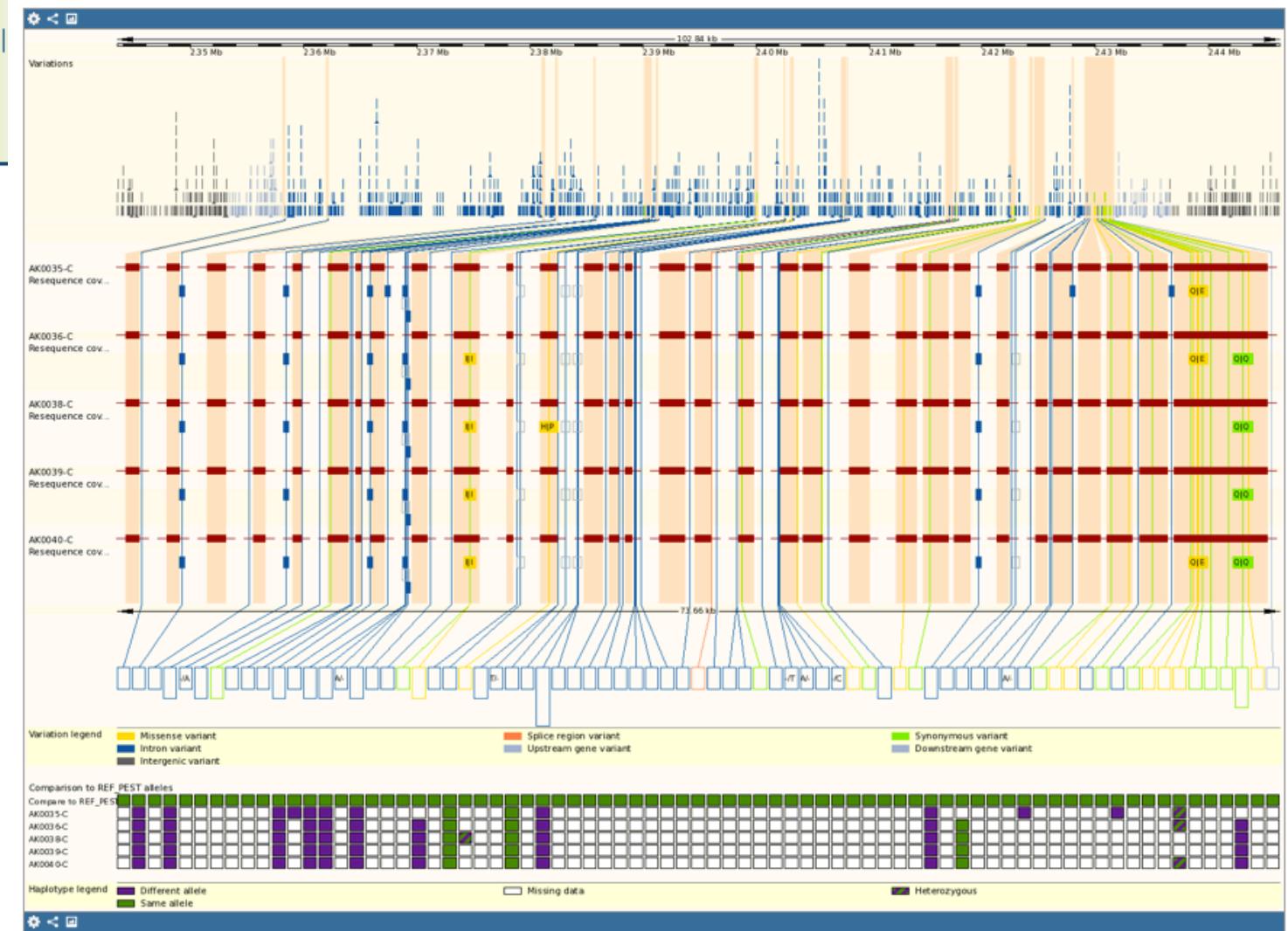
Location: Chromosome 2L:2422652 (forward strand) | [View in location tab](#)

Synonyms: This variation has 2 synonyms - click the plus to show

HGVS names: This variation has 7 HGVS names - click the plus to show

Explore this variation

- Genomic context
- Genes and regulation
- Population genetics
- Individual genotypes
- Linkage disequilibrium
- Phenotype data
- Phylogenetic context
- Flanking sequence



How to contribute?

Population Biology

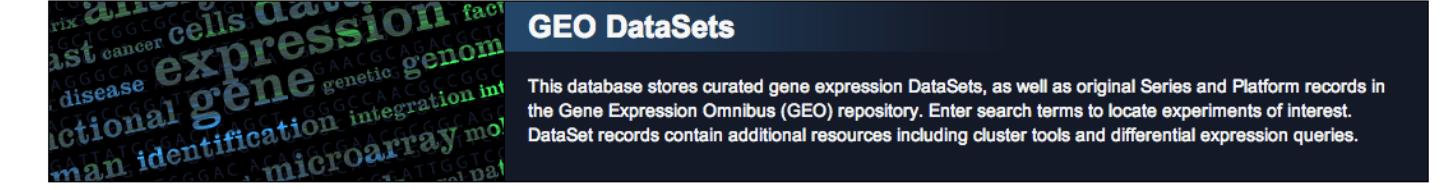
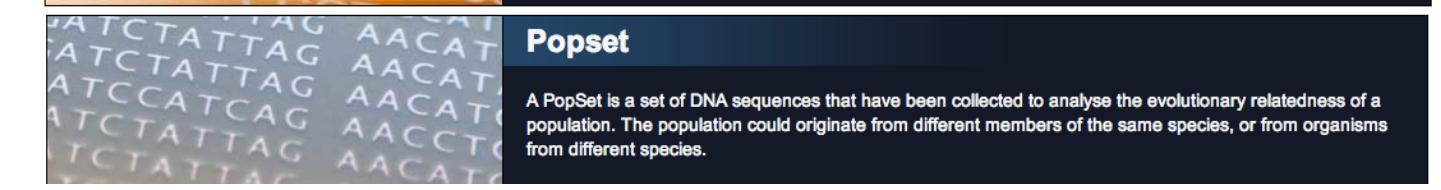
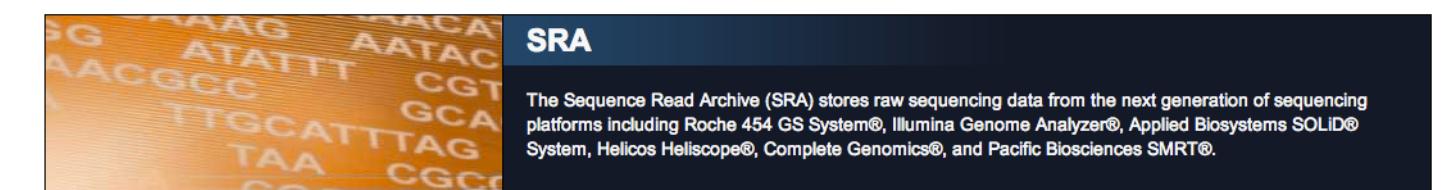
- Submission is via ISA-Tab spreadsheets
- Can share via Google Docs and have real-time 1:1 interaction with a VectorBase curator
- Sample data: ISA-Tab, BioSamples accession
- Insecticide resistance data: ISA-Tab template
- Variation data: SRA accession and VCF

Habtewold Uganda 2013 New ISA-Tab

A	B	C	D	E	F	G	H	I
Source Name	Sample Name	Description	Material Type	Term Source REF	Term Accession Number	Characteristics [sex (EFO:0000695)]	Term Source REF	Term Accession Number
1								
2	Iganga,Uganda H1.F1	Mosquito	individual	EFO	0000542	female	PATO	0000383
3	Iganga,Uganda H2.F1	Mosquito	individual	EFO	0000542	female	PATO	0000383
4	Iganga,Uganda H4.F5	Mosquito	individual	EFO	0000542	female	PATO	0000383
5	Iganga,Uganda H4.HG3	Mosquito	individual	EFO	0000542	female	PATO	0000383
6	Iganga,Uganda H4.F1	Mosquito	individual	EFO	0000542	female	PATO	0000383
7	Iganga,Uganda H4.HG2	Mosquito	individual	EFO	0000542	female	PATO	0000383
8	Iganga,Uganda H4.HG4	Mosquito	individual	EFO	0000542	female	PATO	0000383
9	Iganga,Uganda H4.HG7	Mosquito	individual	EFO	0000542	female	PATO	0000383
10	Iganga,Uganda H5.HG1	Mosquito	individual	EFO	0000542	female	PATO	0000383
11	Iganga,Uganda H5.HG3	Mosquito	individual	EFO	0000542	female	PATO	0000383
12	Iganga,Uganda H5.HG5	Mosquito	individual	EFO	0000542	female	PATO	0000383
13	Iganga,Uganda H5.F1	Mosquito	individual	EFO	0000542	female	PATO	0000383
14	Iganga,Uganda H5.HG2	Mosquito	individual	EFO	0000542	female	PATO	0000383
15	Iganga,Uganda H5.HG4	Mosquito	individual	EFO	0000542	female	PATO	0000383
16	Iganga,Uganda H6.S1	Mosquito	individual	EFO	0000542	female	PATO	0000383

Other data

- Gene symbols, descriptions, citations
- Gene expression (array or sequence based)
- Transcriptomics (RNAseq)
- Proteomics (MS-datasets)



Acknowledgements

Fotis Kafatos, Bob MacCallum, George Christopides, Seth Redmond, Timo Tiirkka

Maggie Werner-Washburne Phil Baker

Kitsos Louis, Pantelis Topalis, Emmanuel Dialynas, Vicky Dritsou



Greg Lanzaro, Yoosook Lee

Charles Taylor



V
EMBL-EBI
Imperial College
NoTre Dame
New MexicO
HaRvard
IMBB
A
Sequencers
Ensembl GEnomes

Daniel Lawson, Gareth Maslen, Mikkel Christensen, Nick Langridge, Derek Wilson, Gautier Koscielny, Karyn Megy, Martin Hammond, Daniel Hughes, Ewan Birney, Paul Kersey

Frank Collins, Greg Madey, Rob Bruggner, Nate Konopinski, EO Stinson, Scott Emrich, Andrew Sheehan, Rory Carmichael, Dave Cieslak, Dave Campbell, Ryan Butler, Katie Cybulski, Neil Lobo, Gloria Calderon, Greg Davis, Antelmo Aguilar, Caleb Reinking,

Bill Gelbart, Susan Russo, Dave Emmert, Pinglei Zhou, Lynn Crosby, Kathy Campbell

TIGR/JCVI, WashU, Broad Institute, Baylor College, Sanger Institute

Martin Donnelly, Dave Weetman, Craig Wilding

Dominic Kwiatkowski, James Stalker

How to stay informed or contact us



VectorBase



@VectorBase



news@vectorbase.org



info@vectorbase.org

