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

Dan Lawson on behalf of VectorBase
EMBL-EBI

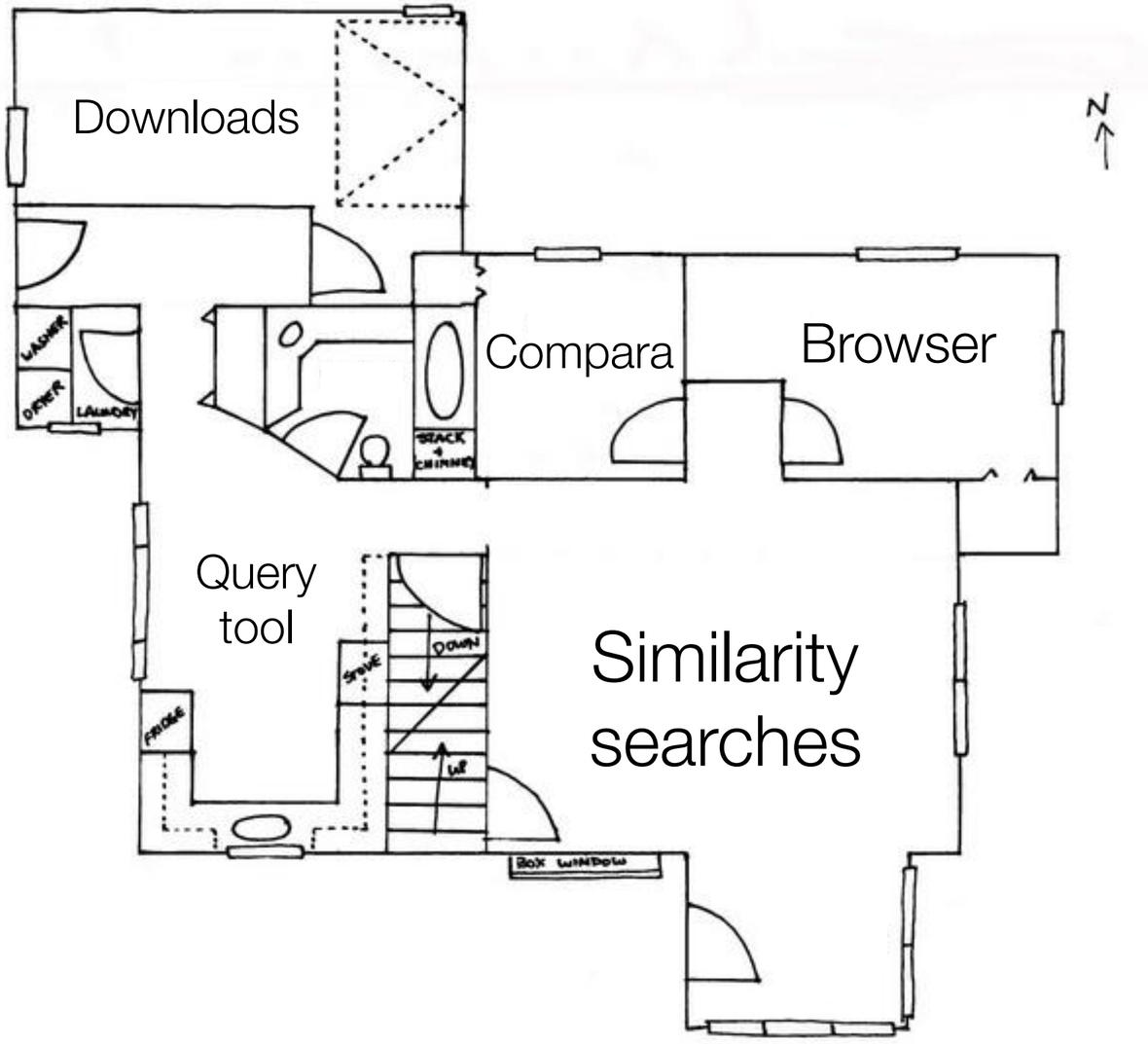
EMBL-EBI

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)

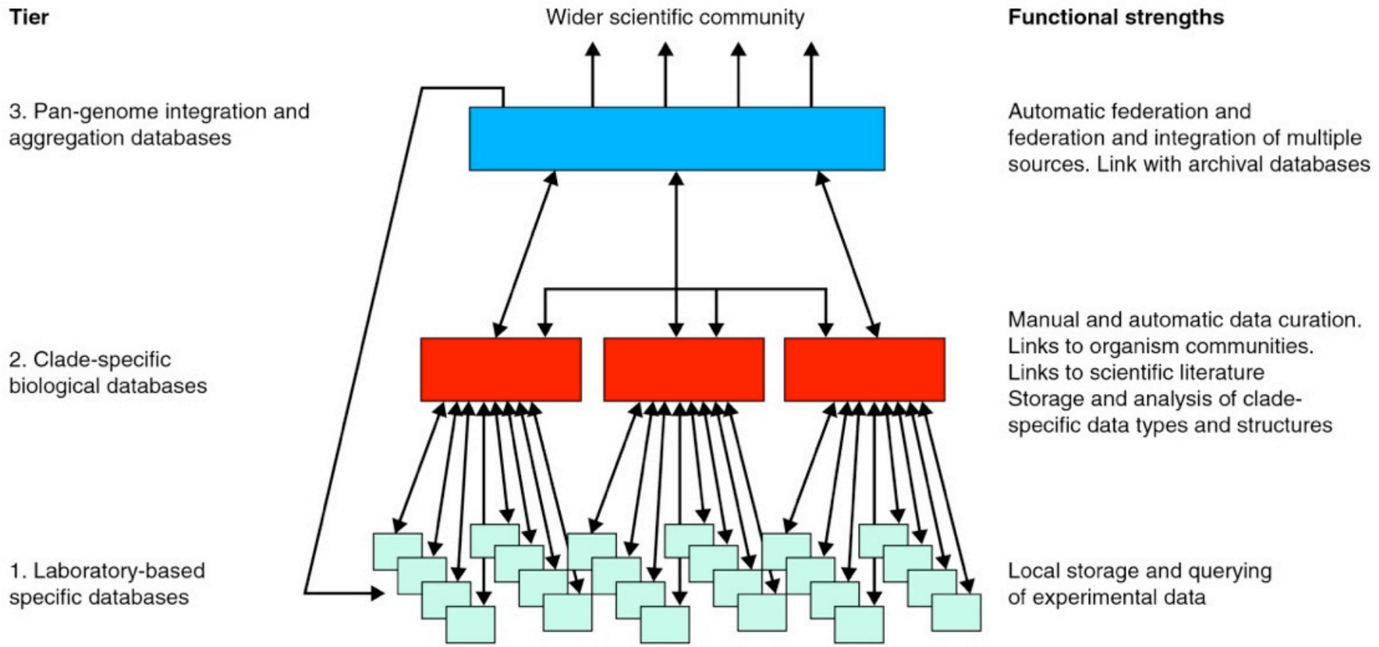
Anatomy of a home

- 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
FlyBase
Ensembl Genomes

Tier 3

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

Finding a home for your data

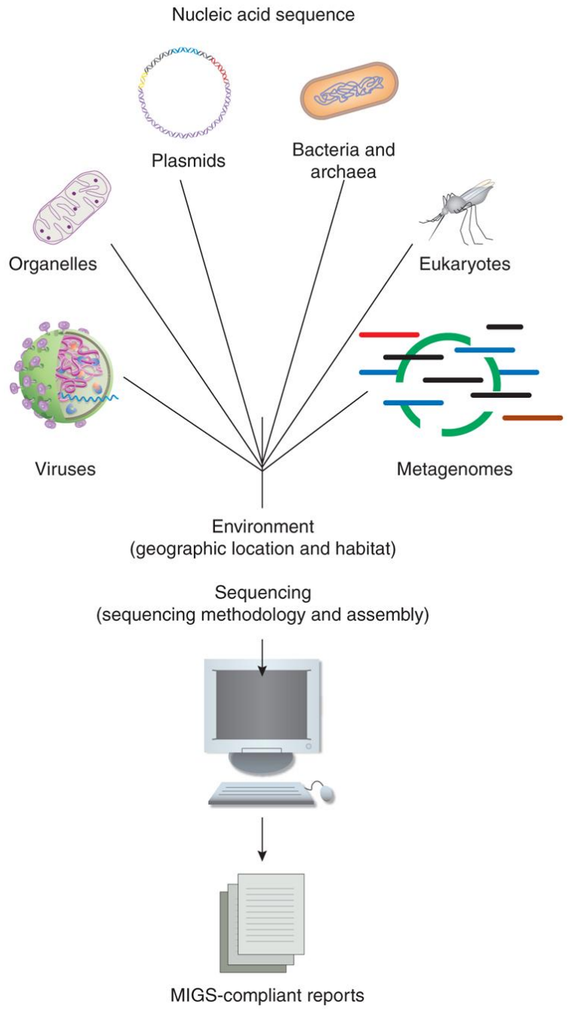
- Visability
  - Integration with the widest possible community
  - Cross-references back to your resource
- Longevity
  - Funding for archival databases is always going to be more secure than your database
- Accreditation
- Publication
  - Many funders and journals require submission prior to publication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

but by far the most important thing is

Submission to the public archival databases

Submission makes you do a number of things

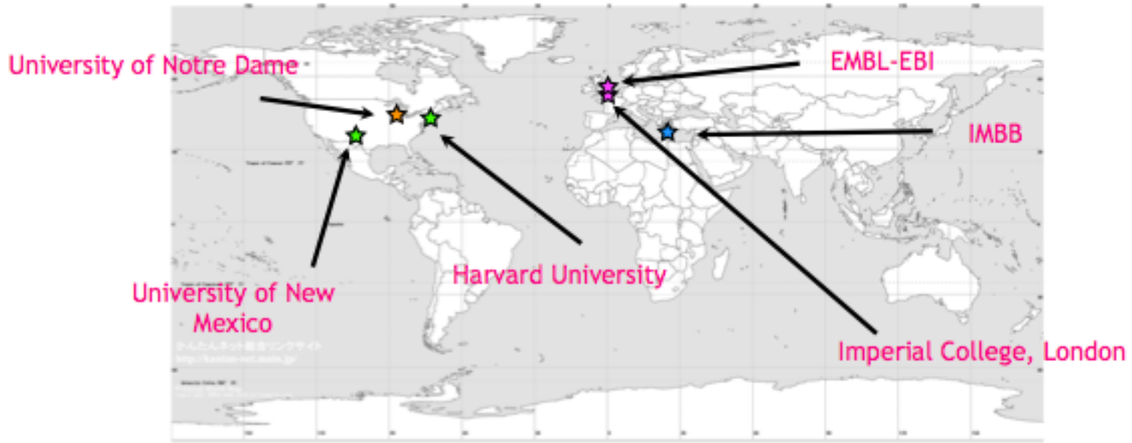
- Requirement to conform to standards
  - Some are mandatory, some advisory
  - Opportunity to capture metadata
    - Minimum information about a genome sequence (MIGS)
    - Metadata gets lost very quickly...
- Encourages good practice
- Explicit nomenclature and versioning



VectorBase: a tier 2 database

VectorBase
Bioinformatics Resource for
Invertebrate Vectors of Human Pathogens

- 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)

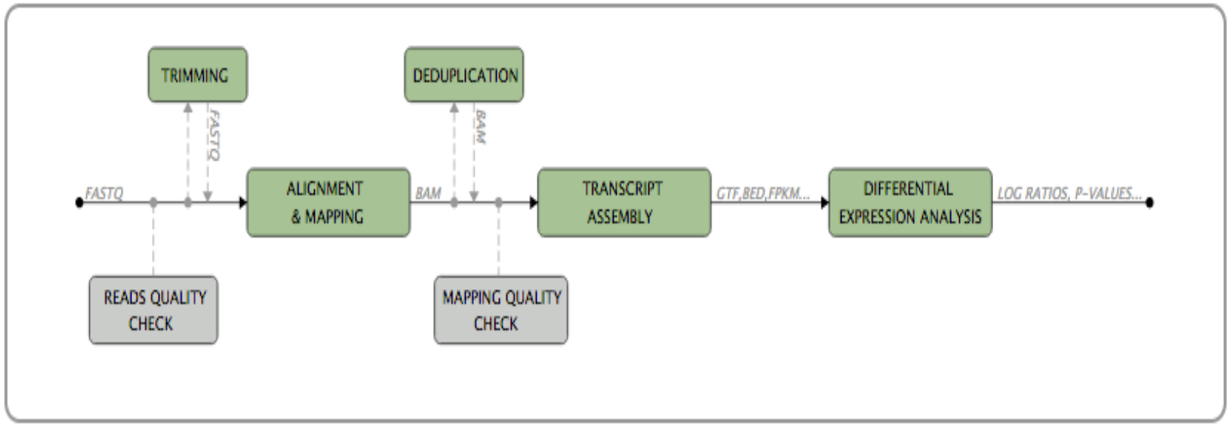
VectorBase (www.vectorbase.org)

- 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

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

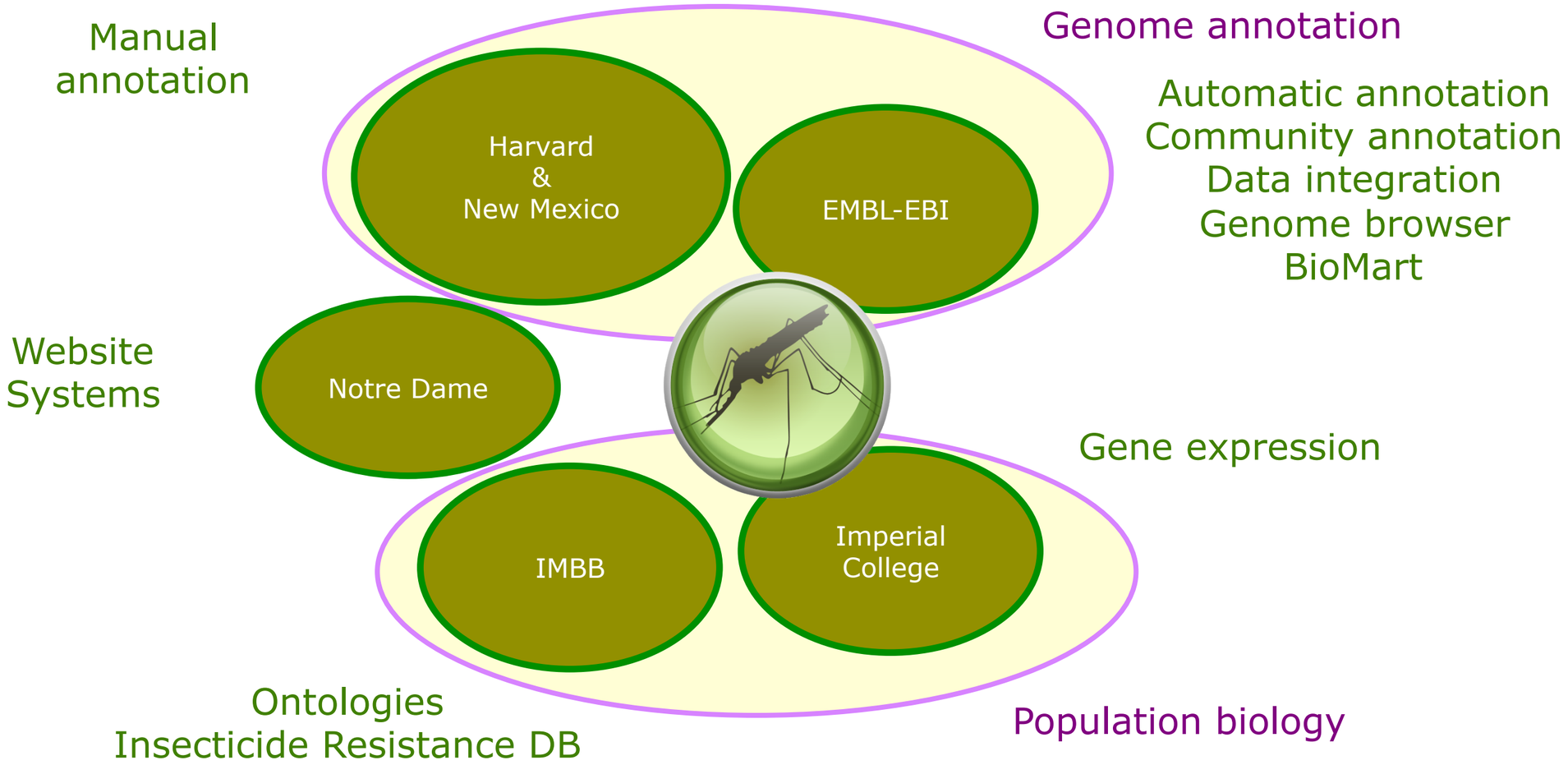


Galaxy

GMOD

http://rnaseq.pathogenportal.org/

VectorBase project roles

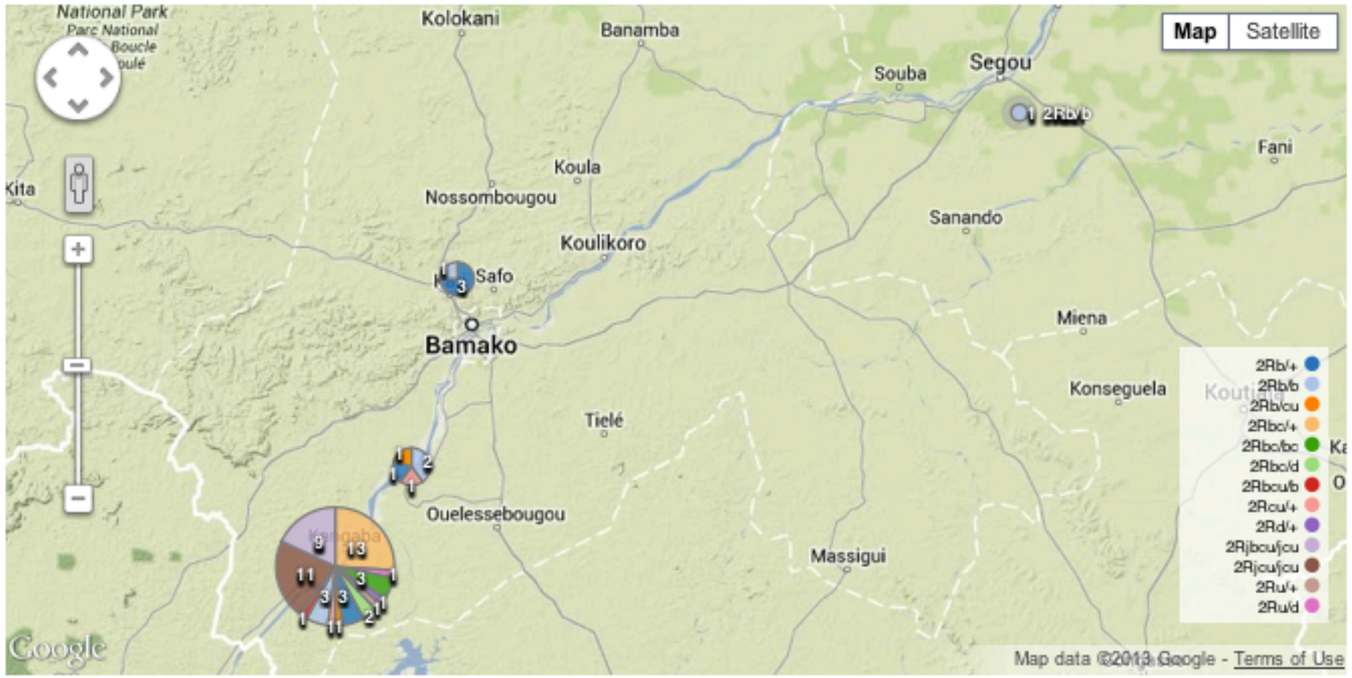


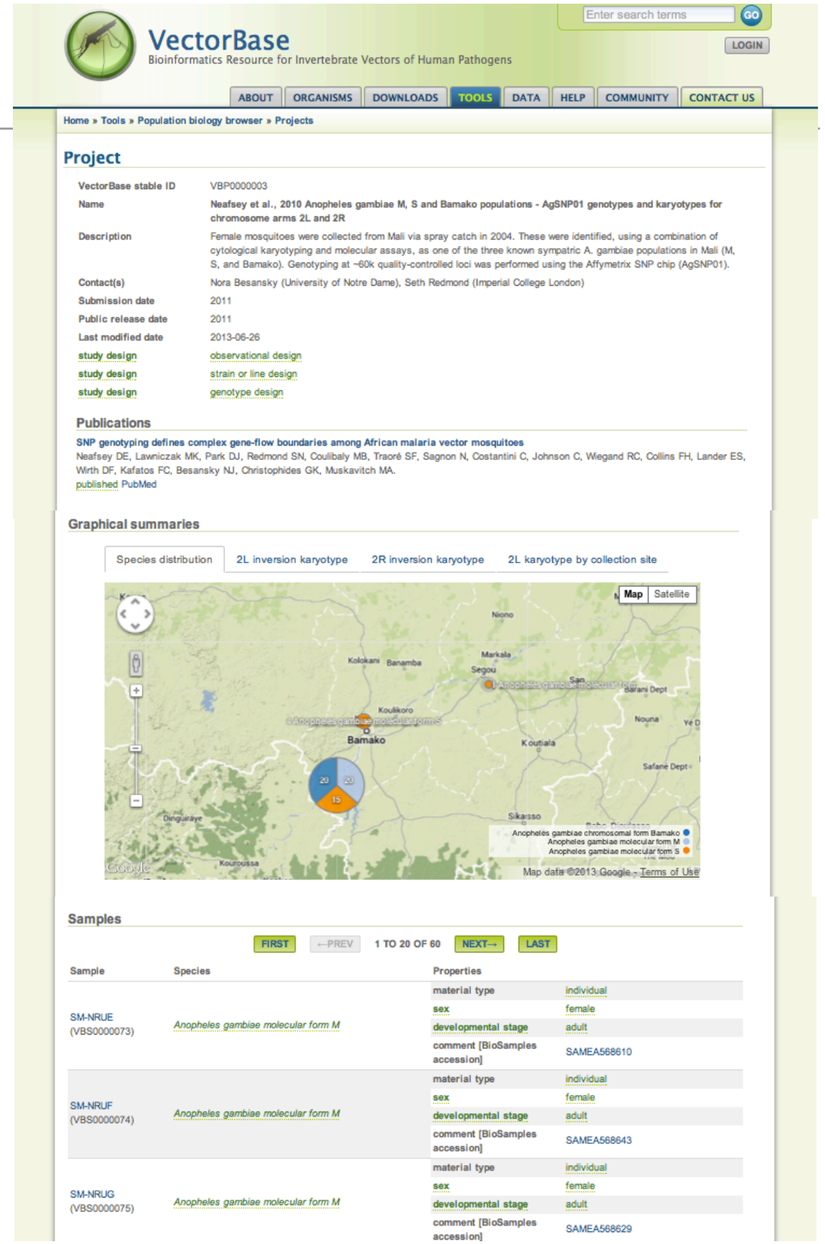
Population biology resource

- 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

Example project: Species and karyotype data

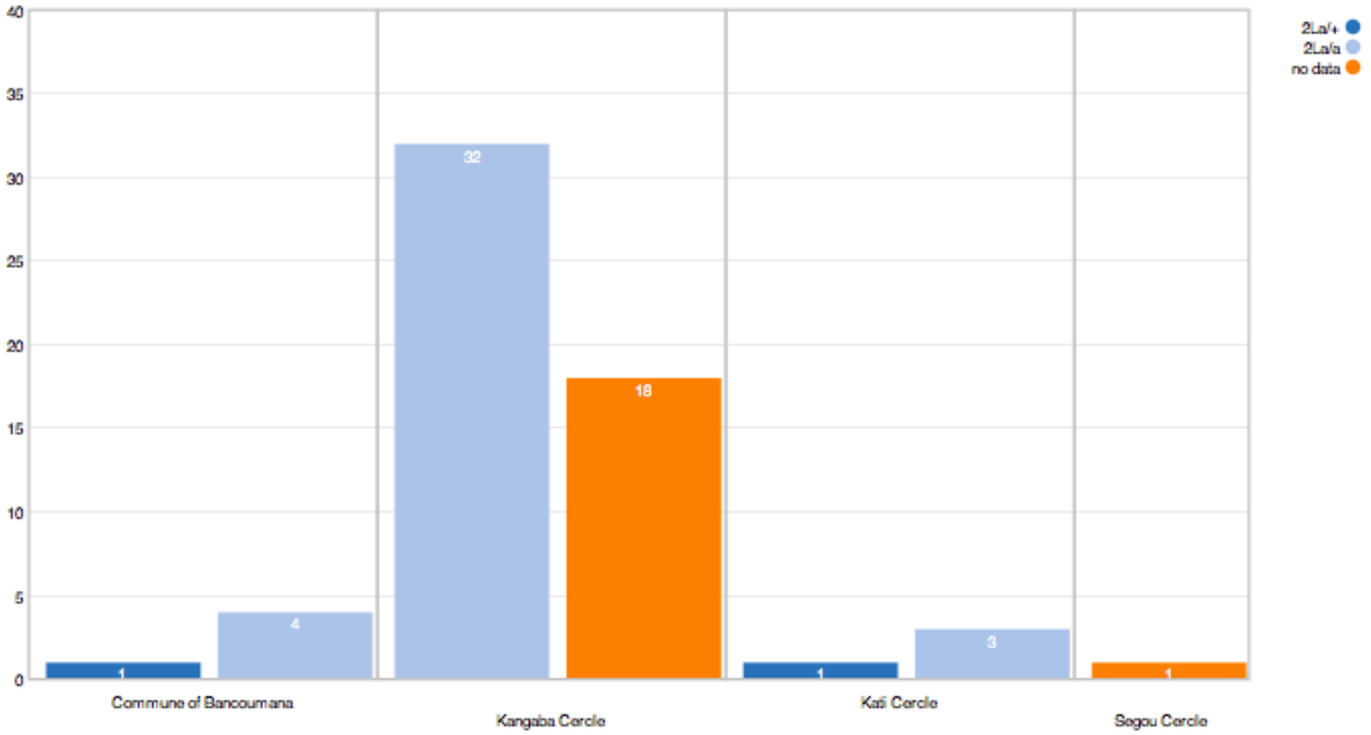
- Project metadata
- Graphical summaries of data (maps, charts)
- Sample information with links to Biosamples (where available)

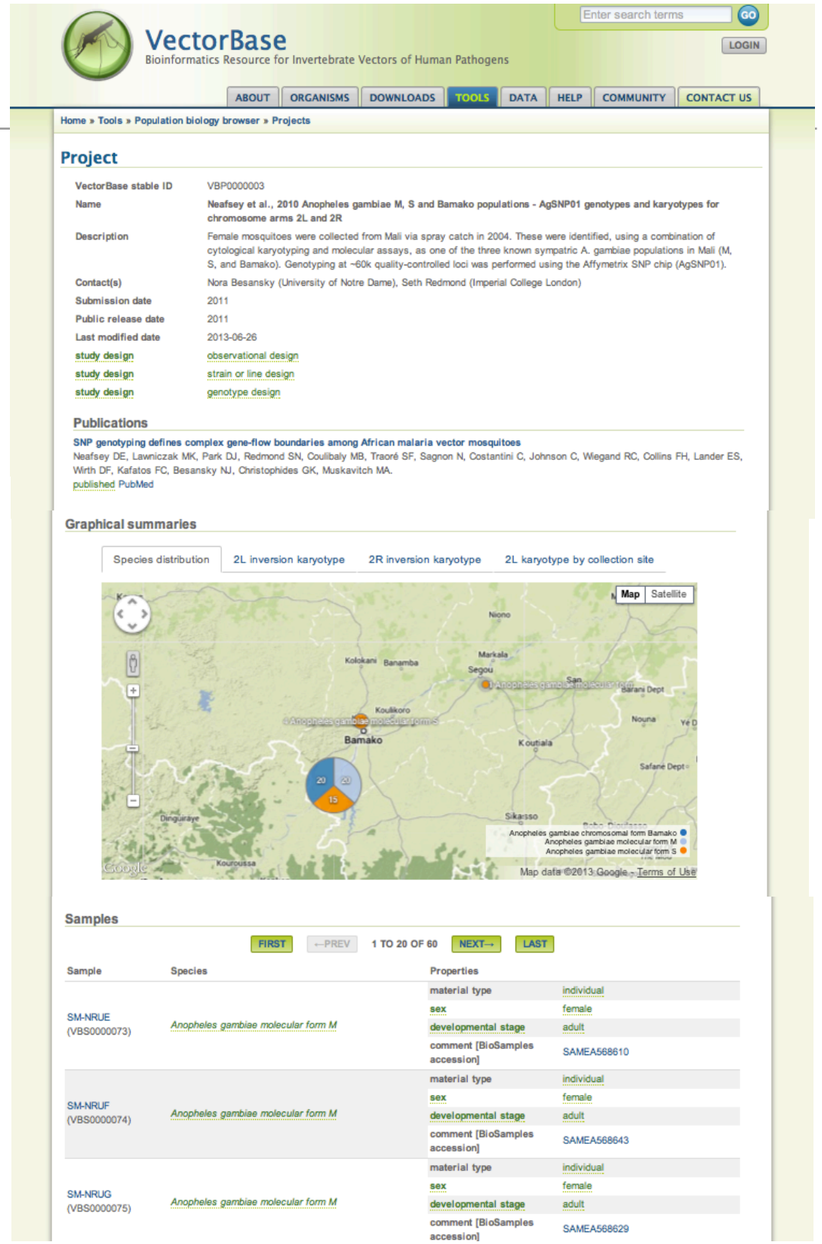




Example project: Species and karyotype data

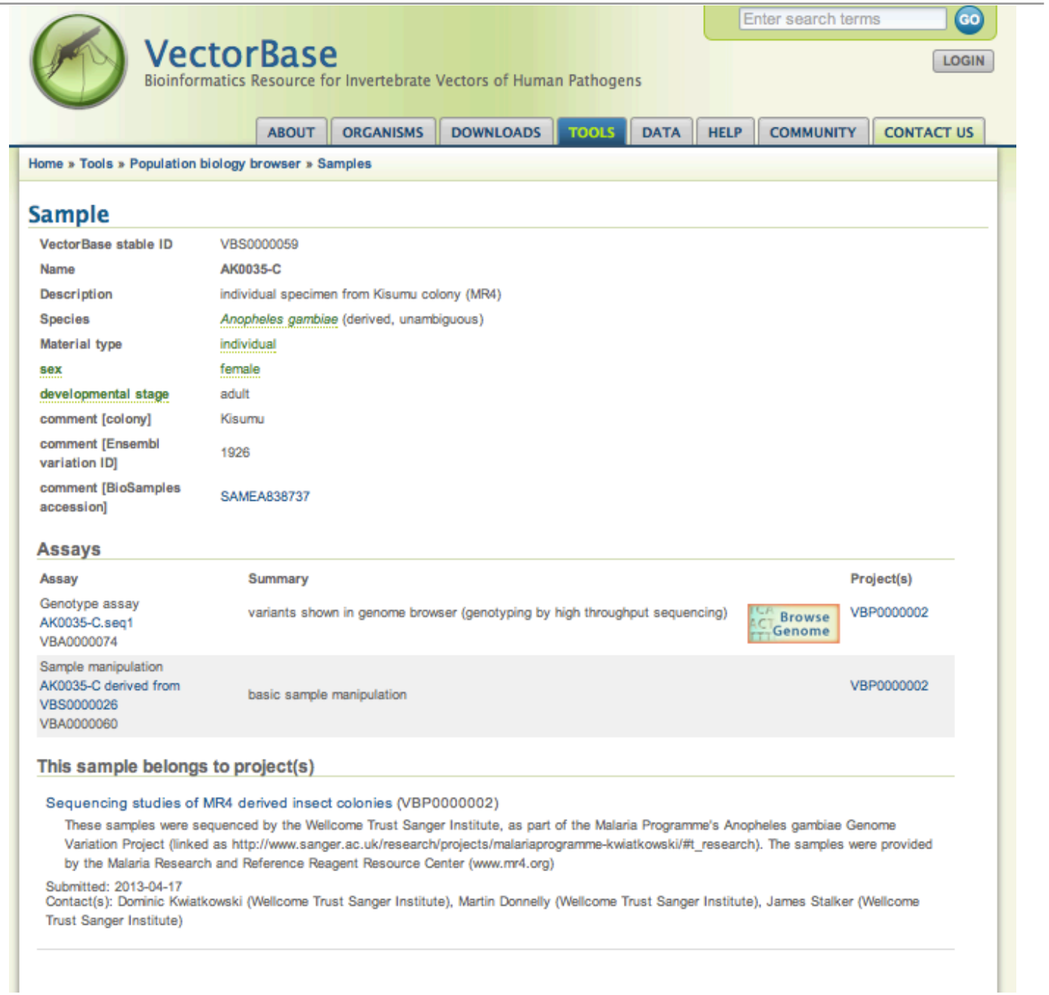
- Project metadata
- Graphical summaries of data (maps, charts)
- Sample information with links to BioSamples (where available)





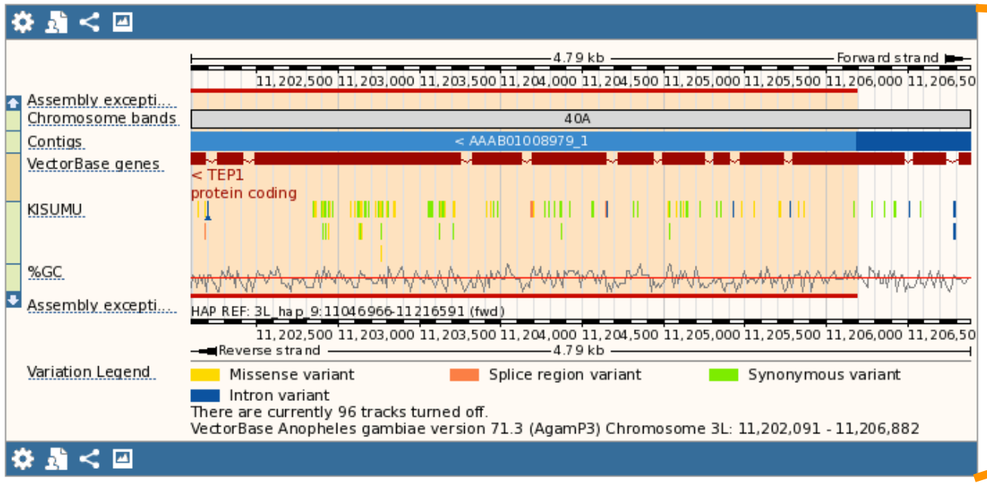
Example sample: NGS re-sequencing & SNP calling

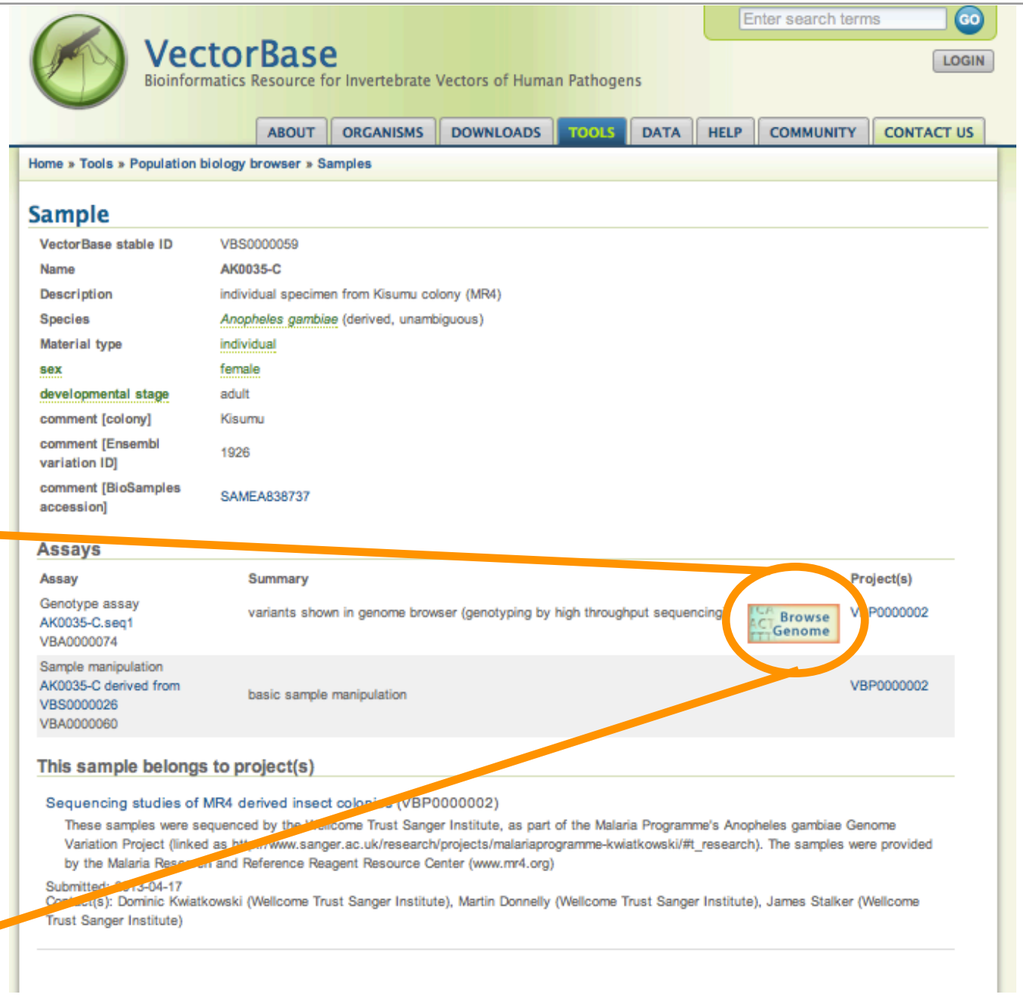
- Sample metadata
- Annotated using ontologies (both VectorBase generated, e.g. MIRO and external e.g. EFO, PATO)
- Links to parental project
- Links to data (BioSamples/SRA)



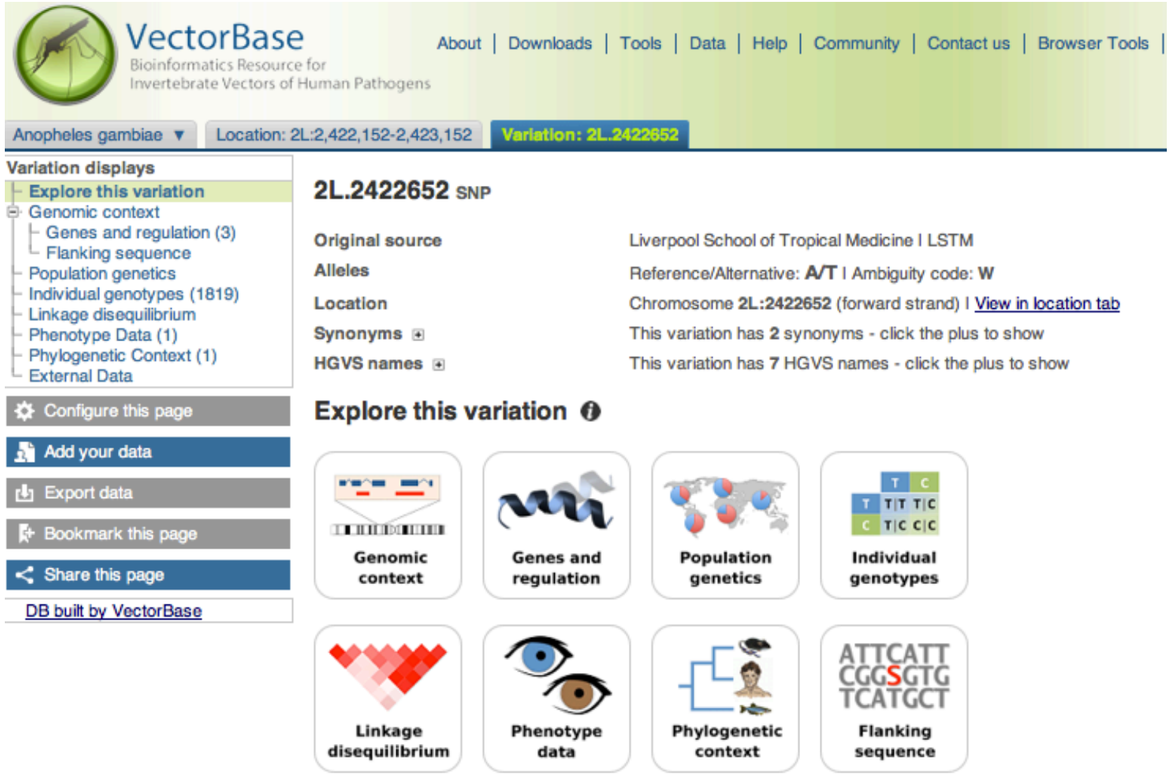
Example sample: NGS re-sequencing & SNP calling

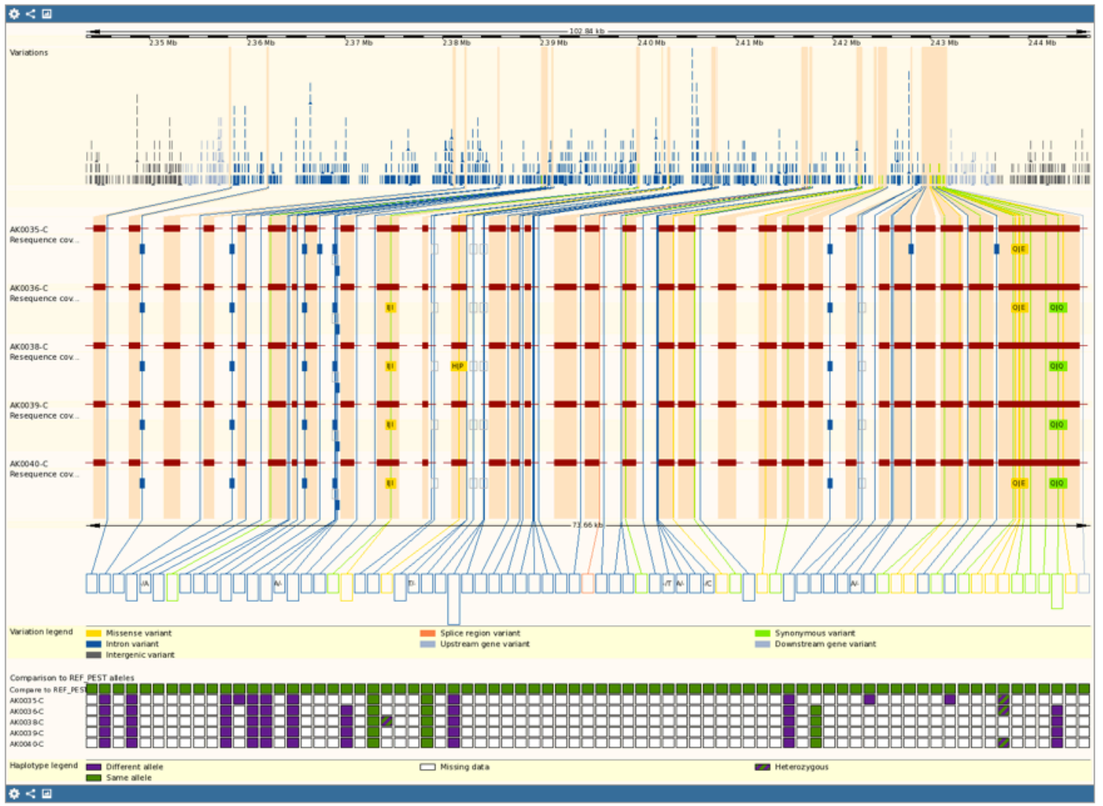
- Sample metadata
- Annotated using ontologies (both VectorBase generated, e.g. MIRO and external e.g. EFO, PATO)
- Links to parental project
- Links to data (BioSamples/SRA)





Population genetics visualisation





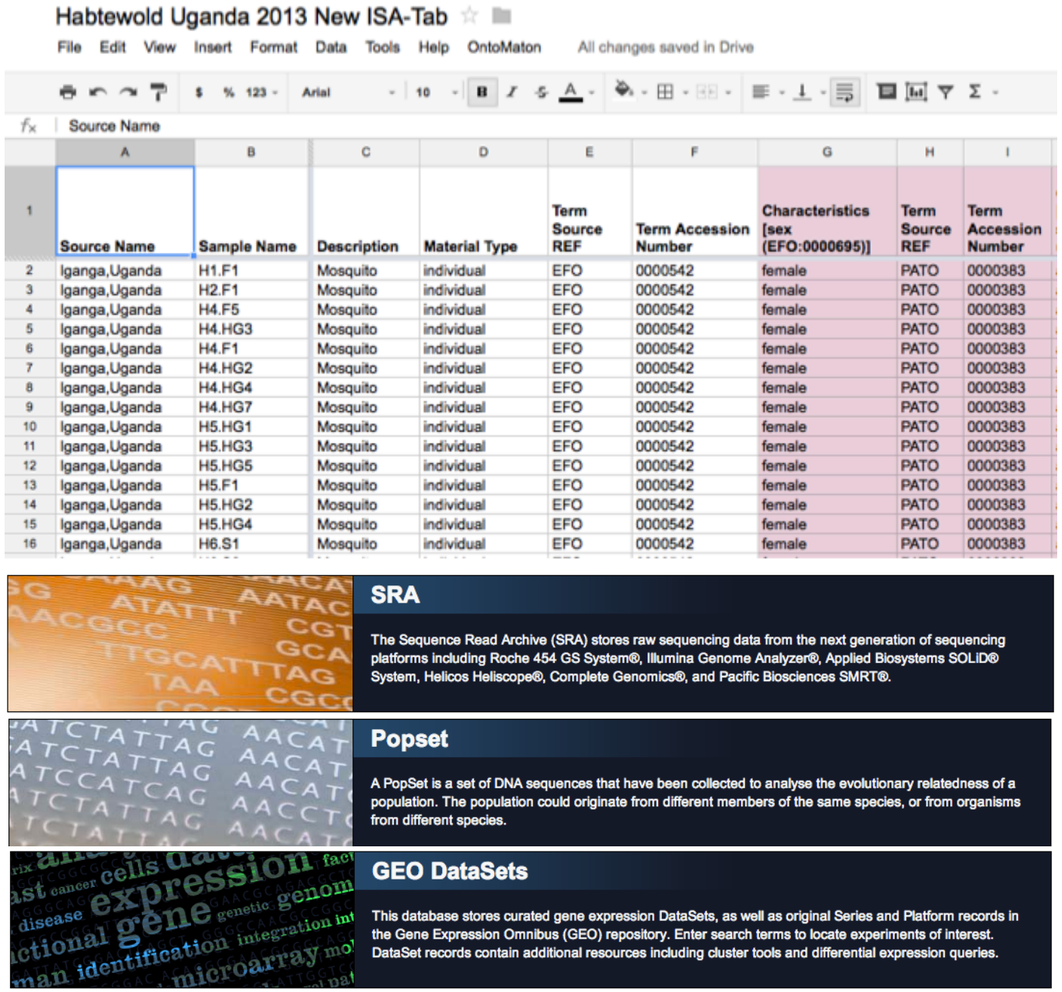
How to contribute?

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

Other data

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



Habtewold Uganda 2013 New ISA-Tab

File Edit View Insert Format Data Tools Help OntoMaton All changes saved in Drive

Source Name

	A	B	C	D	E	F	G	H	I
1	Source Name	Sample Name	Description	Material Type	Term Source REF	Term Accession Number	Characteristics [sex (EFO:0000695)]	Term Source REF	Term Accession Number
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

SRA

The Sequence Read Archive (SRA) stores raw sequencing data from the next generation of sequencing platforms including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD® System, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.

Popset

A PopSet is a set of DNA sequences that have been collected to analyse the evolutionary relatedness of a population. The population could originate from different members of the same species, or from organisms from different species.

GEO DataSets

This database stores curated gene expression DataSets, as well as original Series and Platform records in the Gene Expression Omnibus (GEO) repository. Enter search terms to locate experiments of interest. DataSet records contain additional resources including cluster tools and differential expression queries.

Acknowledgements

Fotis Kafatos, Bob MacCallum, George Christophides, Seth Redmond, Timo Tiirikka

Maggie Werner-Washburne Phil Baker

Kitsos Louis, Pantelis Topalis, Emmanuel Dialynas, Vicky Dritsou

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

UC DAVIS UNIVERSITY OF CALIFORNIA — Greg Lanzaro, Yoosook Lee

UCLA — Charles Taylor

LSTM LIVERPOOL SCHOOL OF TROPICAL MEDICINE — Martin Donnelly, Dave Weetman, Craig Wilding

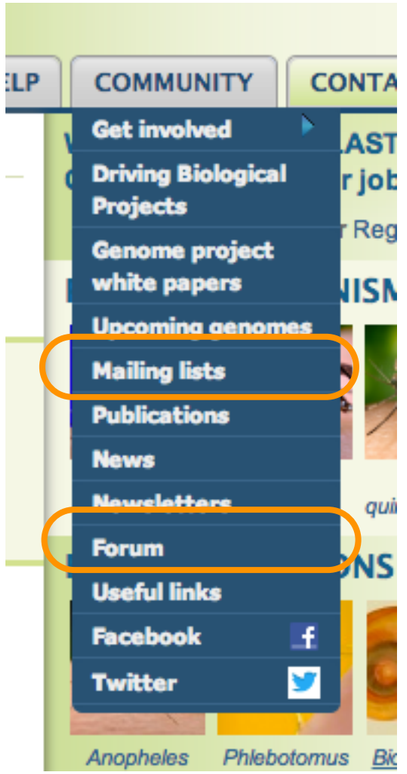
wellcome trust sanger institute — Dominic Kwiatkowski, James Stalker

How to stay informed or contact us

VectorBase

@VectorBase

news@vectorbase.org



info@vectorbase.org