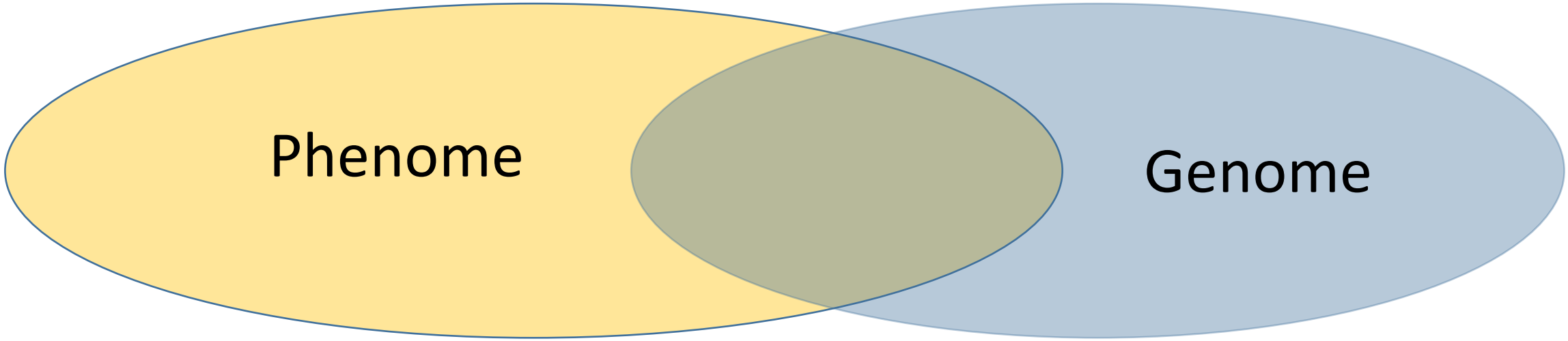
Linking genomic to specimen & species trait data



Specimens

Environment

Population genomics

Phylogenomics

Episode 1: a winge…

Episode 2: Something cool…

With thanks to Paula Mabee!

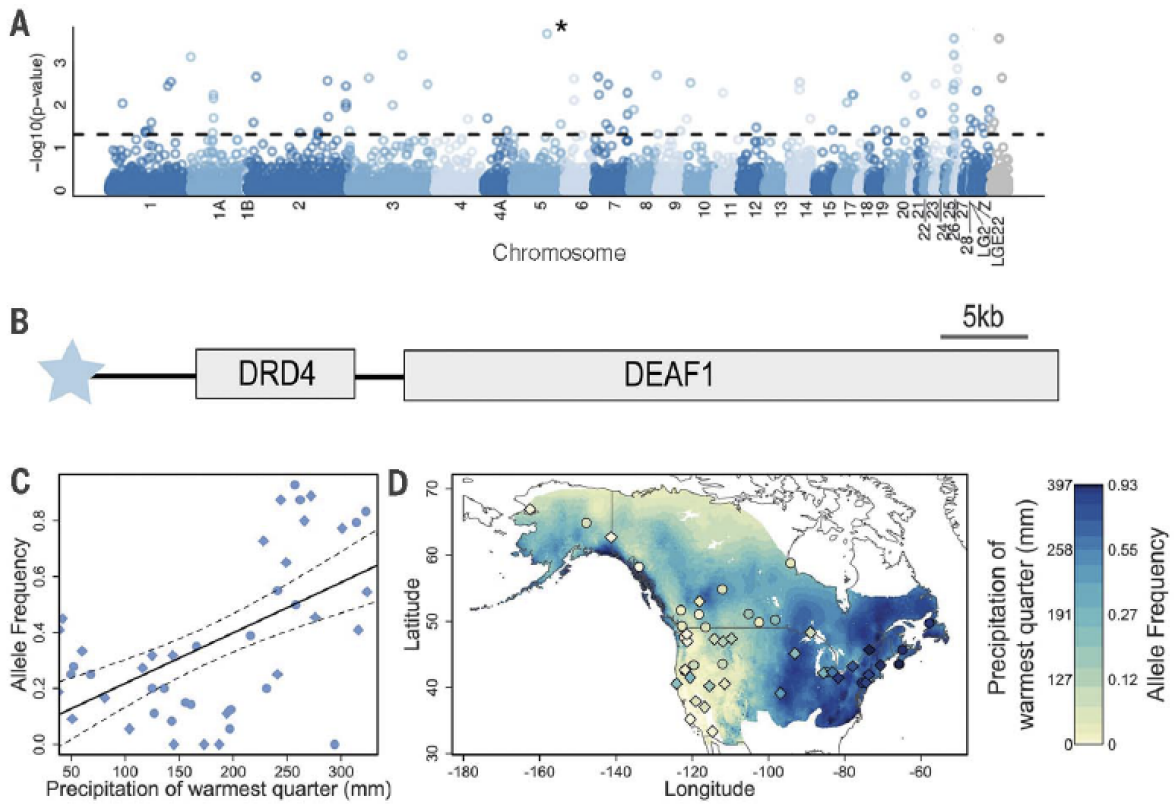
POPULATION GENOMICS (WGS, SNPS)

For many non-model species:

- **SNP screens via RAD or similar, often with no reference genome**
- **GWAS used to detect outlier SNPs correlated to traits or environment**
- **SNP data (here RADseq, ~100K SNPs) available as raw data on ncbi short read archive or EBI-EVA**
- **Often not connectable to specimen records**

Genomic signals of selection predict climate-driven population declines in a migratory bird

Rachael A. Bay,[1,2*] Ryan J. Harrigan,[1] Vinh Le Underwood,[1] H. Lisle Gibbs,[3] Thomas B. Smith,[1,4] Kristen Ruegg[1,5]



TARGET CAPTURE SEQUENCE DATA: the bioinformatics grey zone

SNPs eg ddRAD, GbS

Target capture

- UCEs
- Exons
  - Anchored Enrichment
  - Custom exons

RNASeq

Whole Genome Reseqencing

OpenWings

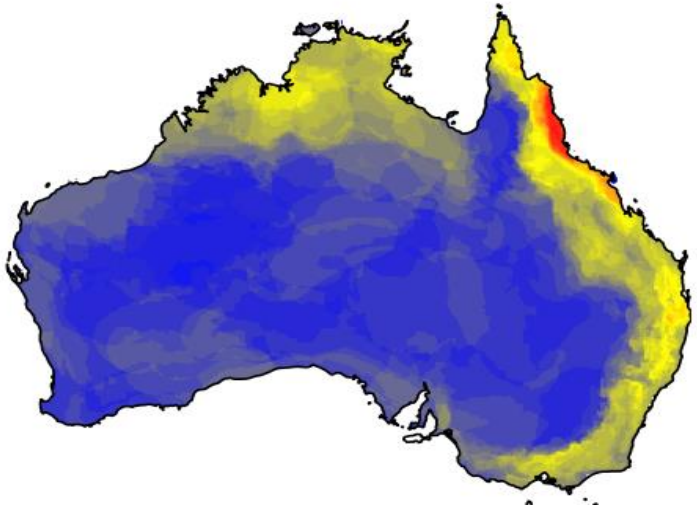
10K avian species, 1000s of UCE loci

OMG
OZ MAMMALS GENOMICS

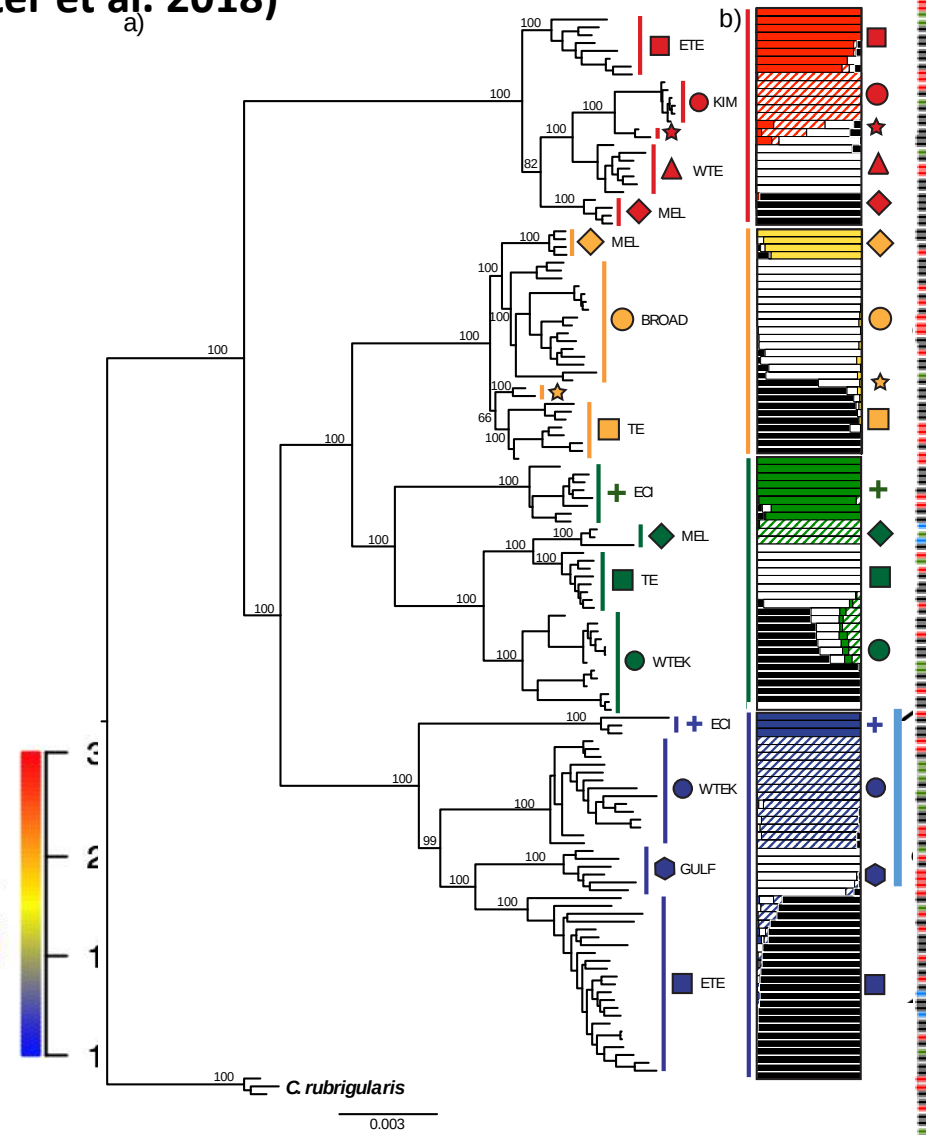
700 species,
>2000 specimens,
~2K exons

Exon capture to link micro to macrtoevolution: Eugongylus skinks (Bragg et asl. 2018, Potter et al. 2018)

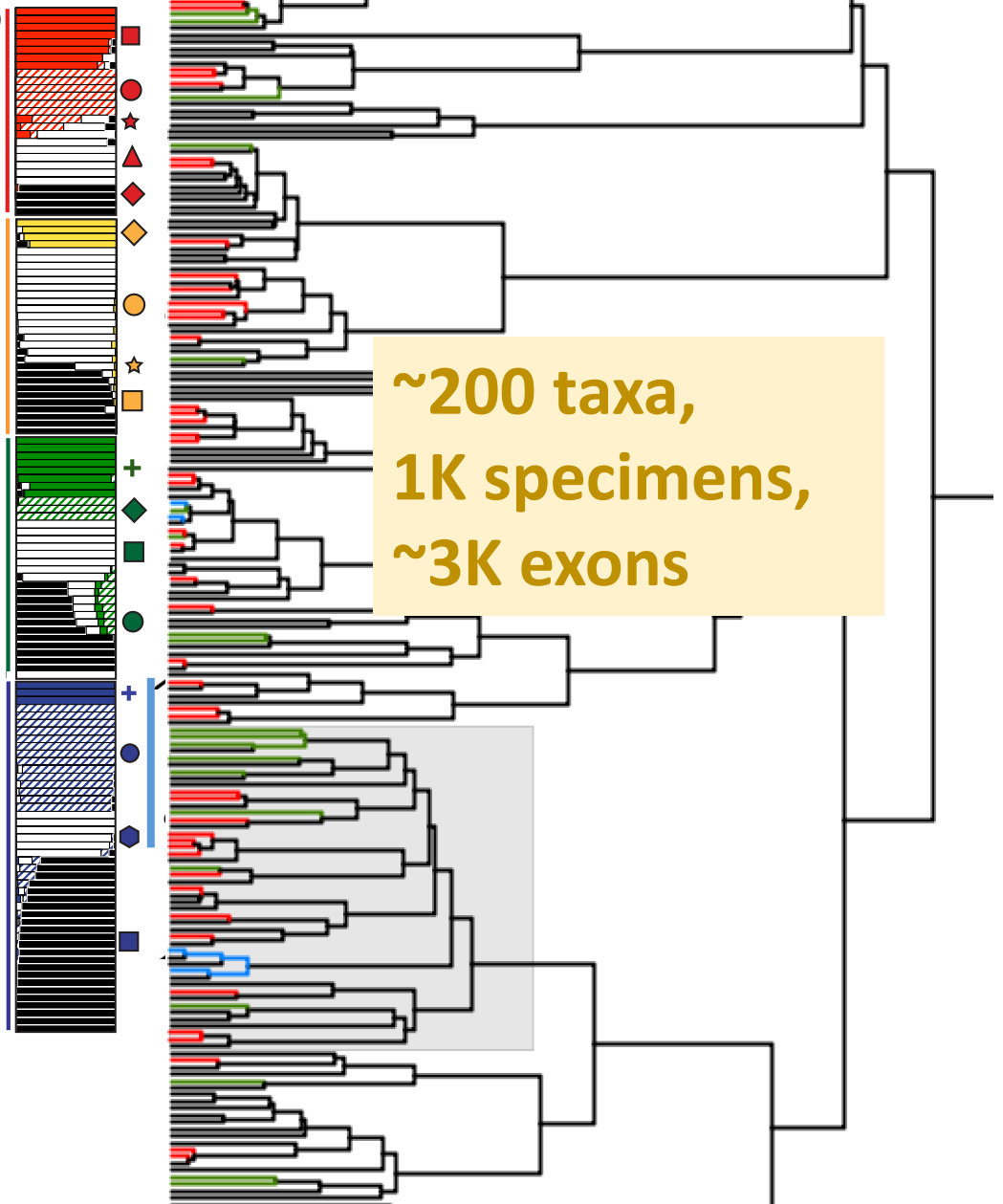
Eugongylus



a)



b)



~200 taxa,
1K specimens,
~3K exons

Phylogenomics of a rapid radiation: the Australian rainbow skinks

Jason G. Bragg[1,2*], Sally Potter[1], Ana C. Afonso Silva[1,3], Conrad J. Hoskin[4], Benjamin Y. H. Bai[1,5] and Craig Moritz[1]

Raw data

Code & alignments

DRYAD About For researchers For organizations

Data from: Phylogenomics of a rapid radiation: the Australian rainbow skinks

Bragg JG, Potter S, Silva ACA, Hoskin CJ, Bai BYH, Moritz C

Date Published: January 17, 2018

DOI: https://doi.org/10.5061/dryad.v1d32

BMC Evolutionary Biology

Eugongylus Accession: PRJNA289283 ID: 289

Eugongylus group skinks targeted loci

Targeted exon capture sequencing of Eugongylus group skinks.

Accession	PRJNA289283
Data Type	Targeted Locus (Loci)
Scope	Multispecies
Organism	**Eugongylus** [Taxonomy ID: 99192] Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Lepidosauria; Squamata; Bifurcata; Unidentata; Scinciformata; Scincidae; Eugongylinae; Eugongylus
Submission	Registration date: 8-Jul-2015 **Australian National University**
Relevance	Evolution

ncbi SRA Bioproject

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
SRA Experiments	240
OTHER DATASETS	
BioSample	342

▾ SRA Data Details

Parameter	Value
Data volume, Gbases	251
Data volume, Tbytes	0.16

Biosample data

biomaterial provider	Western Australian Museum
birth date	missing
birth location	missing
breeding history	missing
cell subtype	missing
cell type	missing
collected by	missing
collection	missing
culture c	missing
death date	missing
disease	missing
disease stage	missing
genotype	missing
geographic location	missing
growth protocol	missing
health state	missing
isolation source	Meentheena, Western Australia, Australia
latitude and longitude	21.4161 S 120.4542 E
phenotype	missing
sample type	missing
specimen voucher	WAMR139409
storage conditions	missing

The Genomic Observatories Metadatabase

https://www.geome-db.org/

The Genomic Observatories Metadatabase (GeOMe): A new repository for field and sampling event metadata associated with genetic samples

John Deck, Michelle R. Gaither, Rodney Ewing, Christopher E. Bird, Neil Davies, Christopher Meyer, Cynthia Riginos, Robert J. Toonen, Eric D. Crandall

Published: August 3, 2017 • https://doi.org/10.1371/journal.pbio.2002925

moorea BIOCODE

TAAAGC TATTAA GAATAC Diversity of the Indo-Pacific Network

Smithsonian Institution

GORDON AND BETTY MOORE FOUNDATION

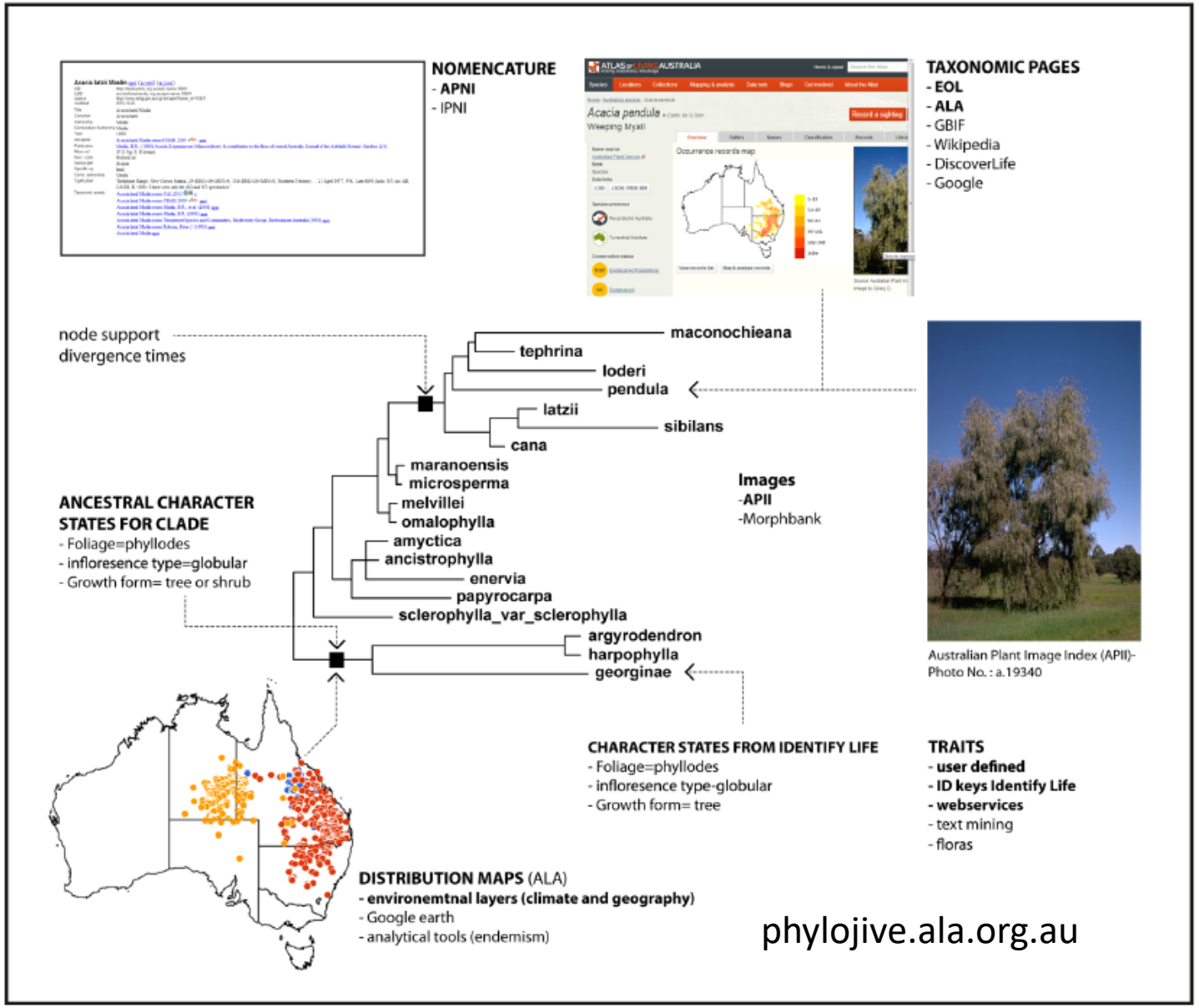
NSF

National Evolutionary Synthesis Center NESCent

ALA Phylojive

Correcting the disconnect between phylogenetics and biodiversity informatics

JOSEPH T. MILLER & GARRY JOLLEY-ROGERS 2014; Zootaxa 3754 :195-200



Only for tips = accepted names

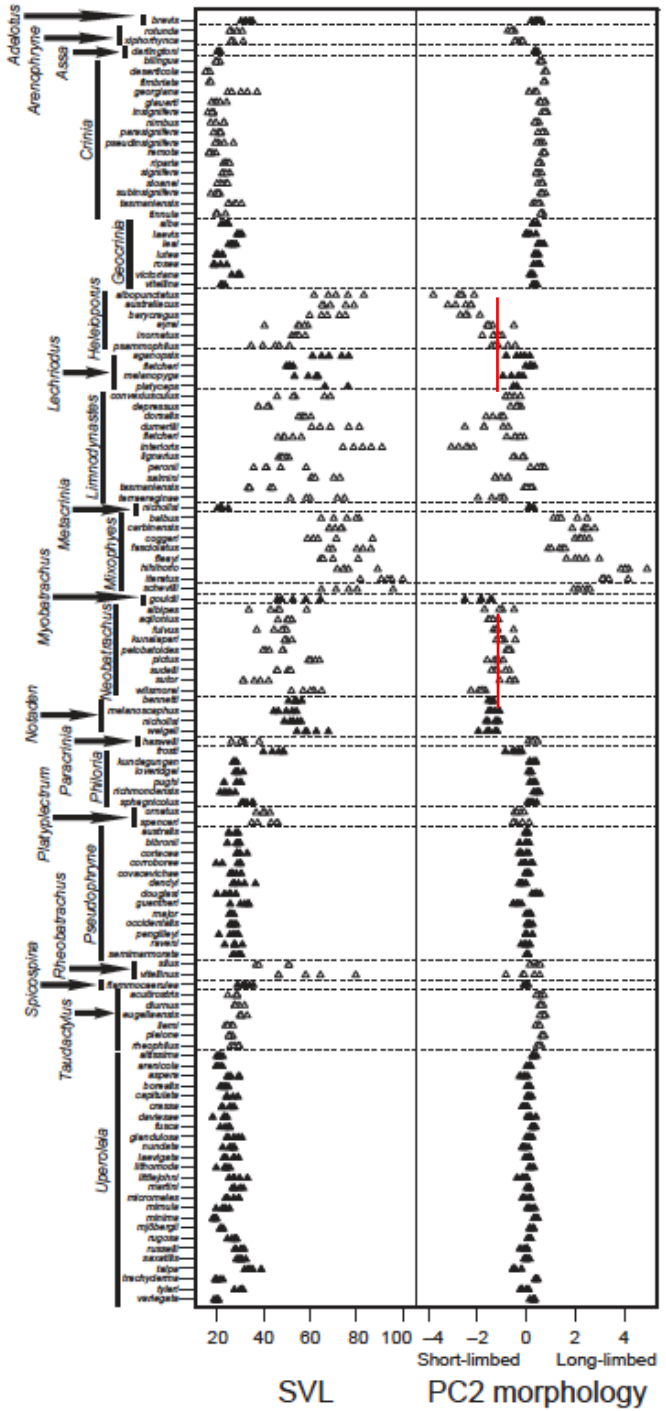
Example:

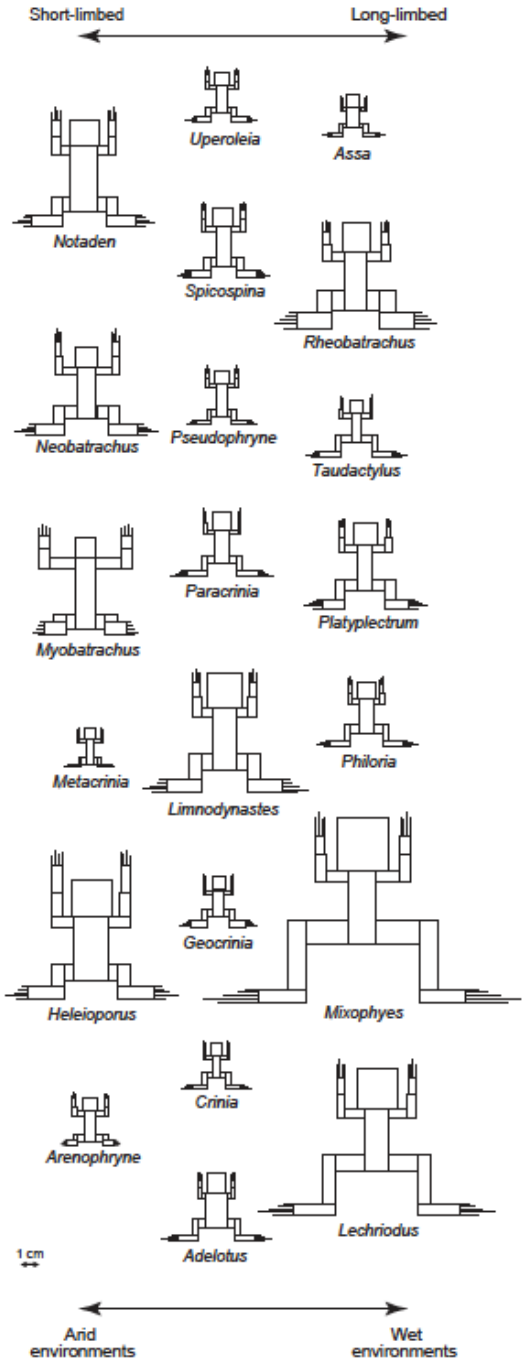
Morphological evolution in myobatrachid frogs

Vidal-Garcia et al. 2014
J. Evol. Biol. 27:181-192

B. Maryan

M. Anstis

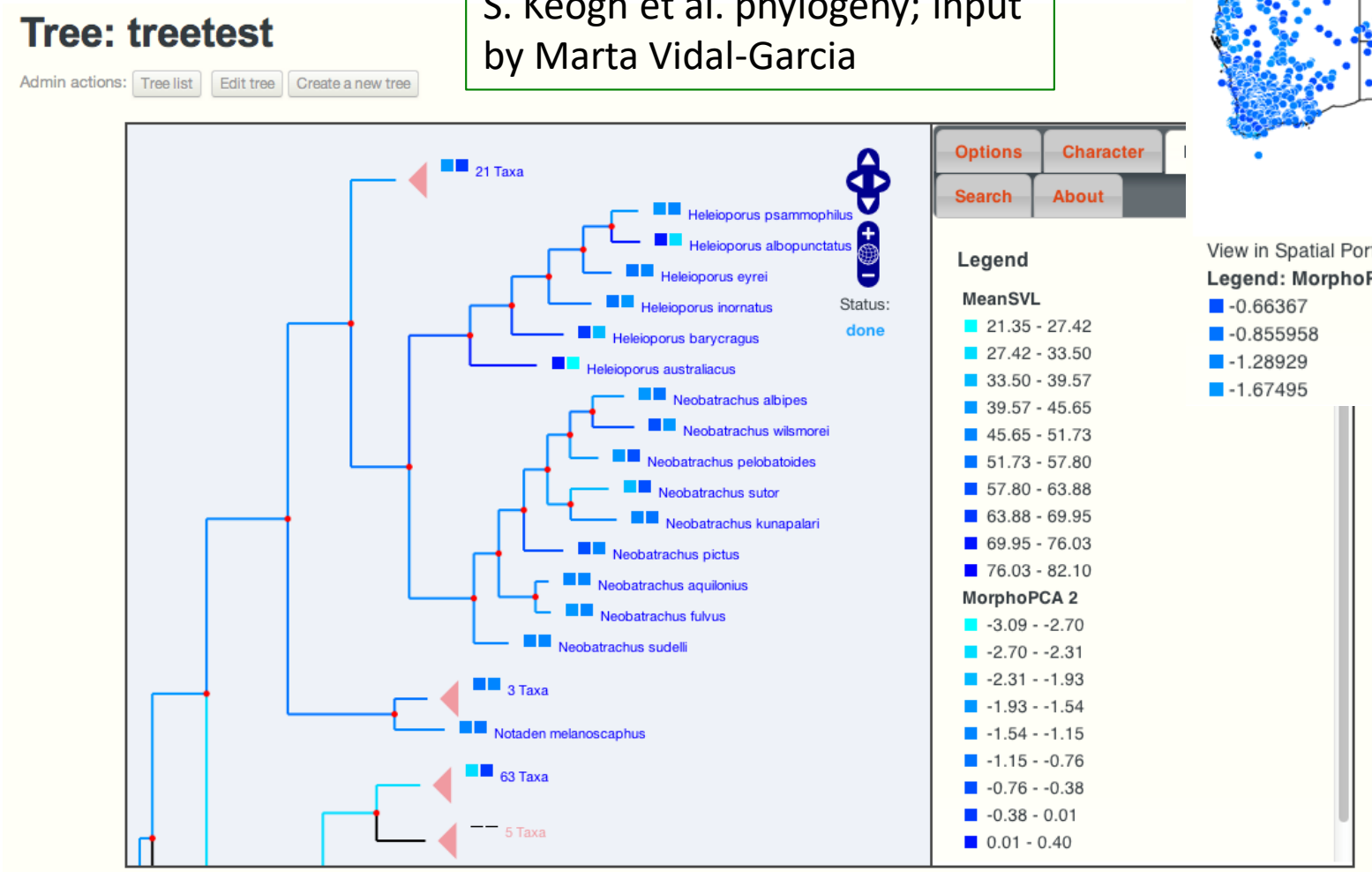


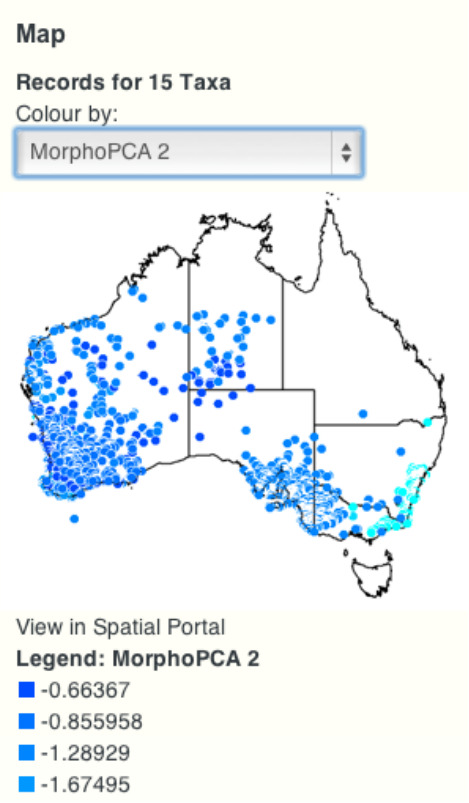


Now in phylojive…

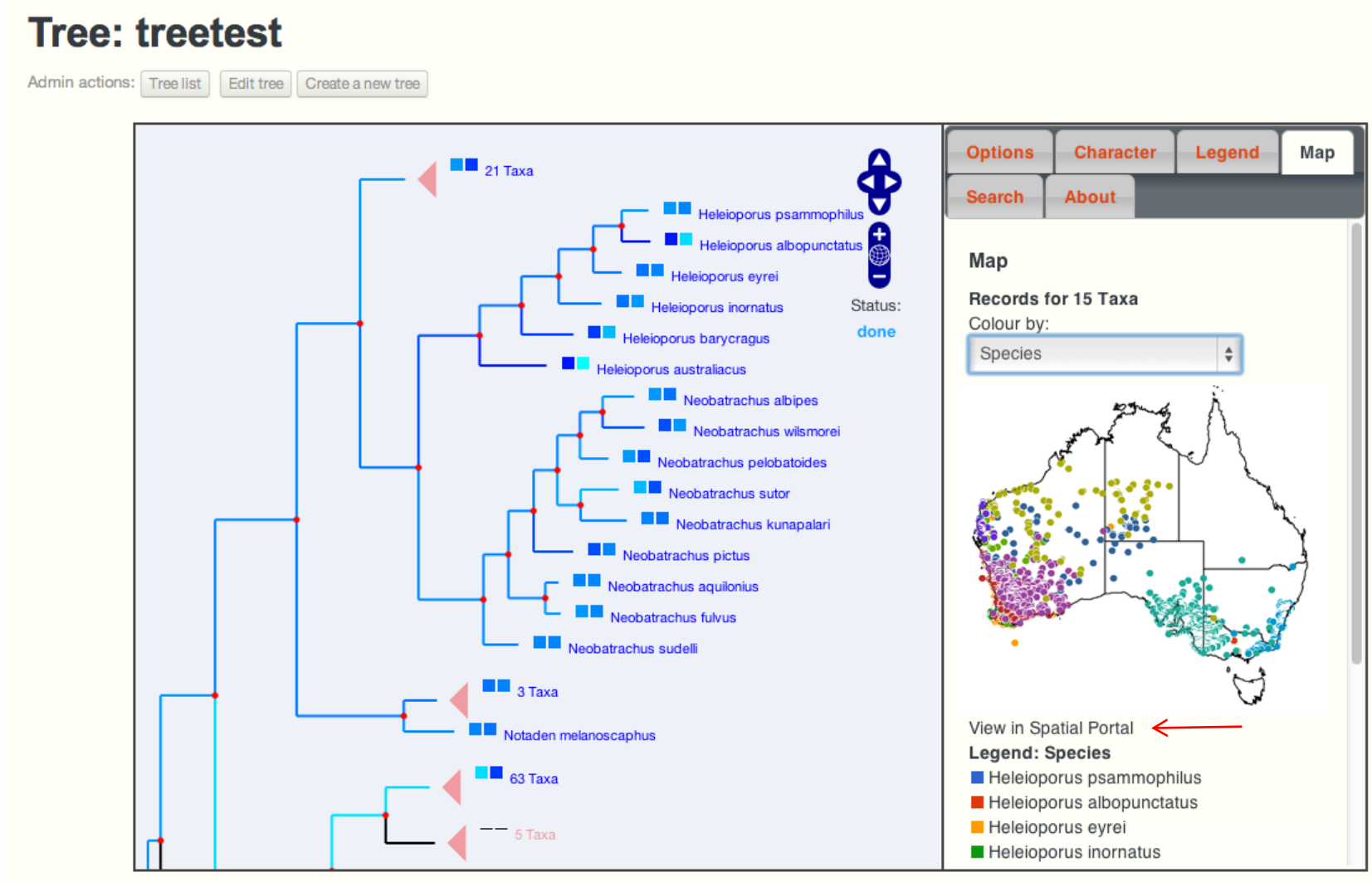
Mapping traits

S. Keogh et al. phylogeny; input by Marta Vidal-Garcia

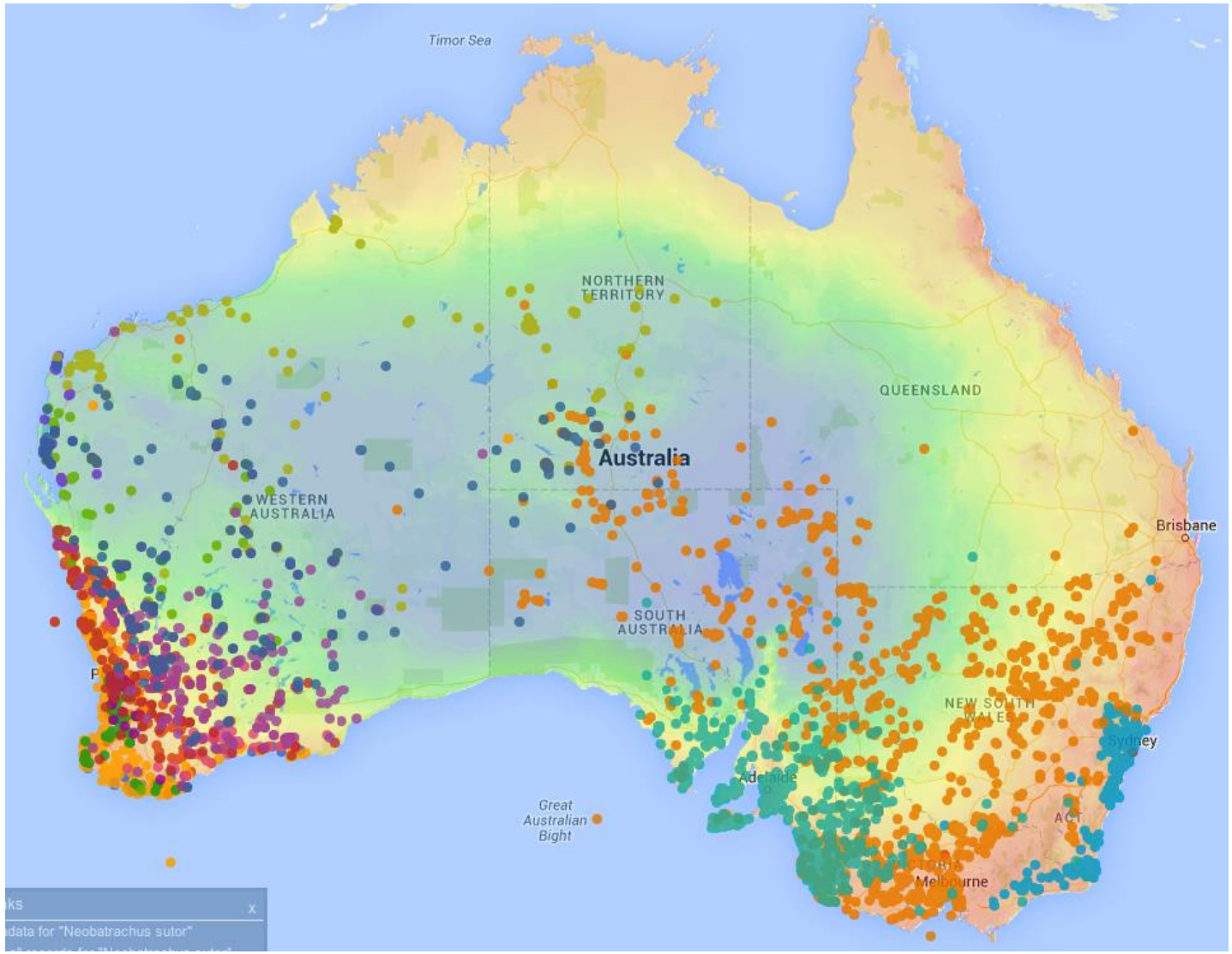




…and Mapping distributions



And now in ALA spatial portal
(on annual mean moisture index)



Reflections on bioinformatics, genomes & traits

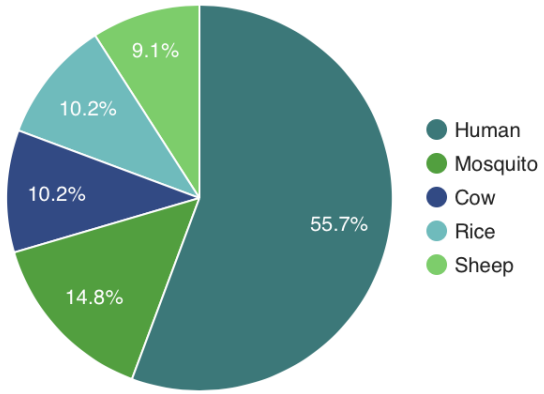
- Capturing and sharing trait data will enrich our understanding of G <-> P in micro and macroevolution
- Comparative WGS is the ideal but still need genome subsampling for most organisms (and research budgets)
- There is a bioinformatics gap connecting museum specimens to subgenome data (esp. target capture)
- Phylogenetic visualization and retrieval of museum data is powerful for exploratory analyses of G + E <-> P

European Variation Archive

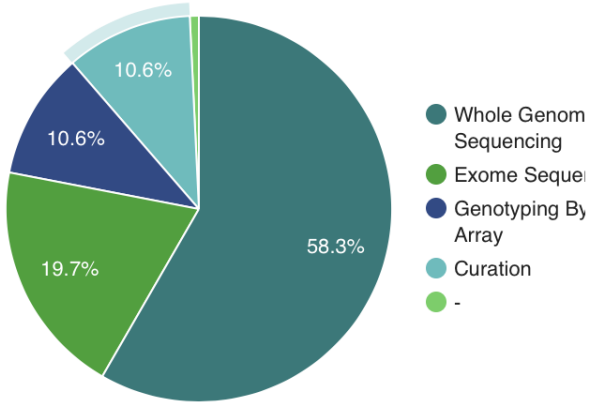
Support for non-human variant data archival and accessioning is transitioning from dbSNP to EVA from September 2017.

Short genetic variants studies (<50bp)

Top 5 Species

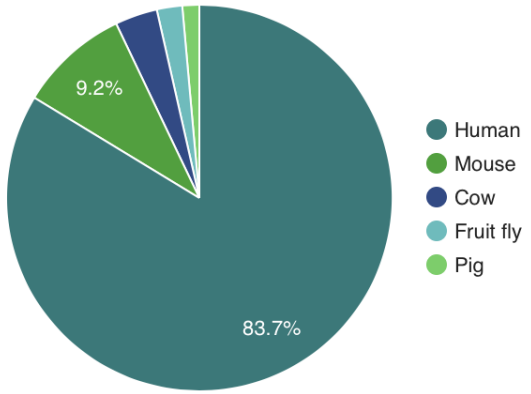


Top 5 Types

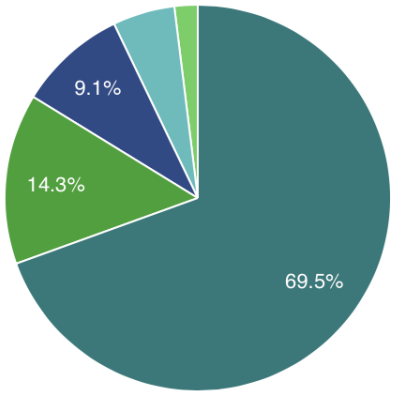


Structural variants studies (>50bp)

Top 5 Species



Top 5 Types

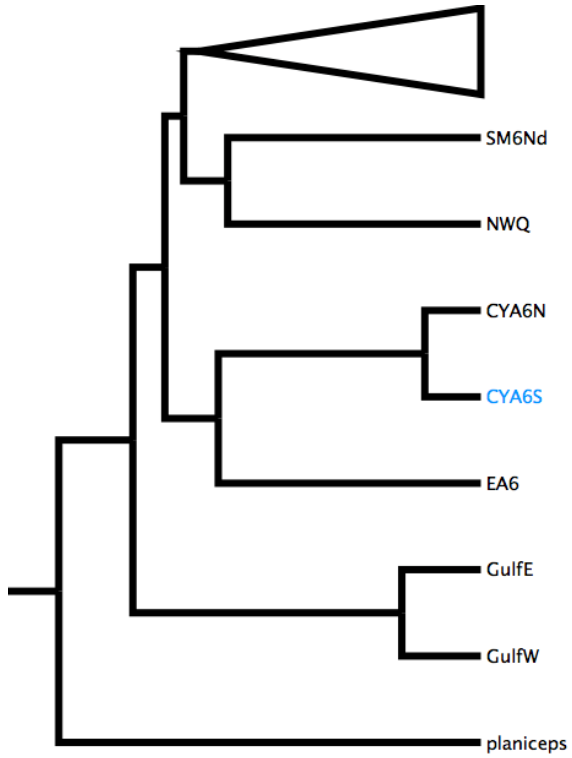


Enabling “phylojive” for phylogeographic lineages

Combine:

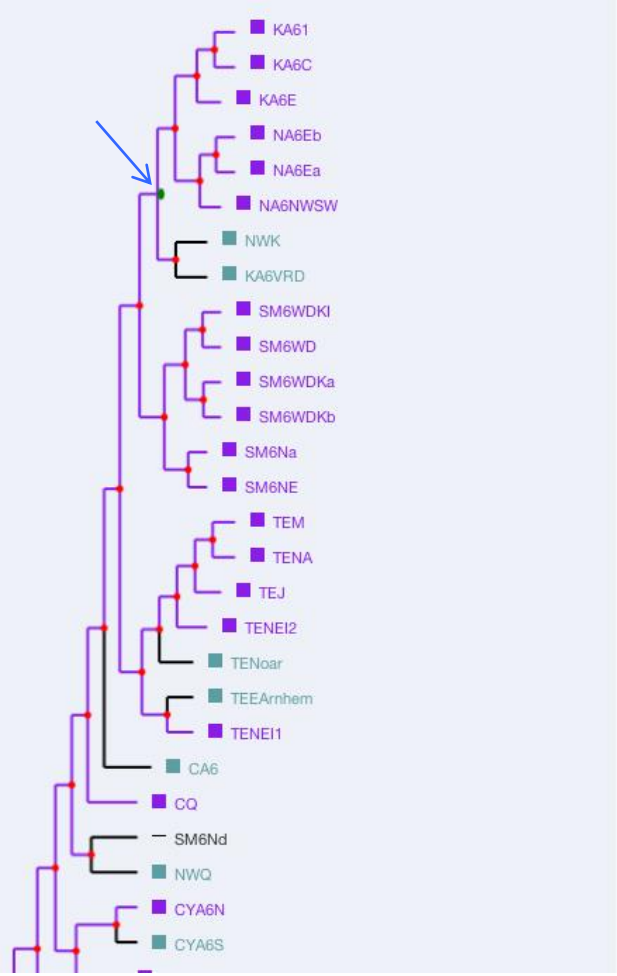
- user-input tree (phylojive) with
- user-input records (sandbox) using informal names

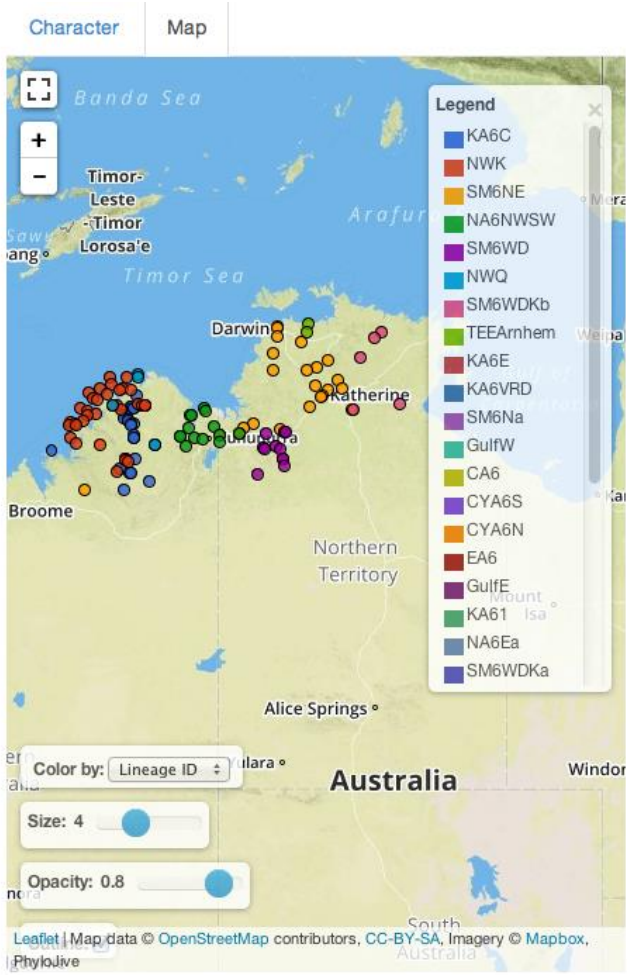
⇒Allow node based visualisation in geographic & environmental space

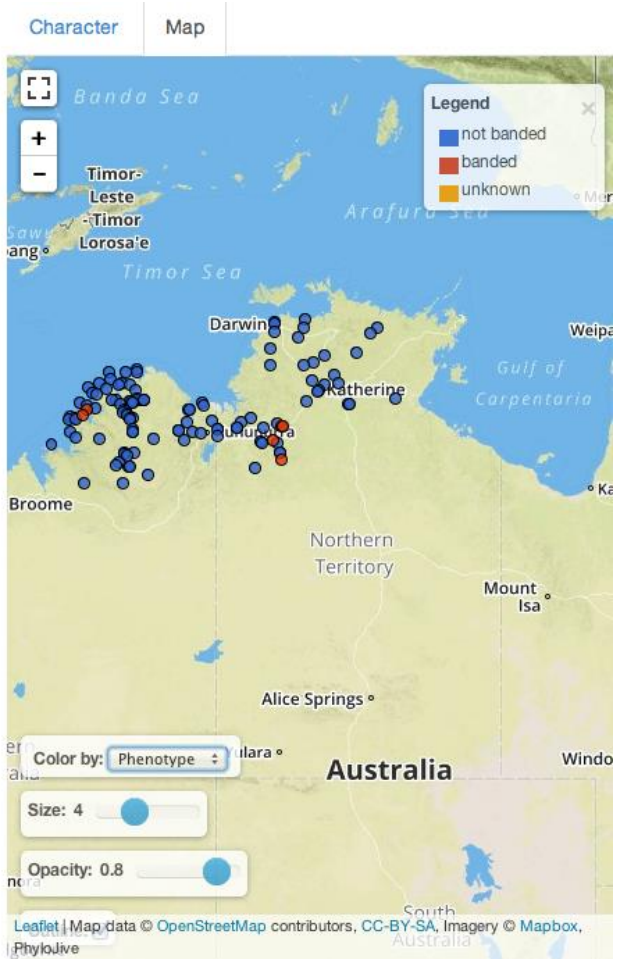


	A	B	C	D	E	F
1	**sample ID**	**lineage ID**	**Location**	**Latitude**	**Longitude**	**phenotype**
2	ABTC77094_	CYA6S	31k N Einasleigh on Mt Surprise I	-18.2945	144.0766	not banded
3	SEW8462_CY	CYA6S	Georgetown-MtGarnet Rd	-18.294454	144.076613	not banded
4	CCM0119_CY	CYA6S	BLAC 37	-19.53237	144.06488	not banded
5	CCM0115_CY	CYA6S	BLAC TIP	-19.5363	144.19704	not banded

Example: *H. binoei* lineages







[Live URL for H. binoei example]

European Variation Archive

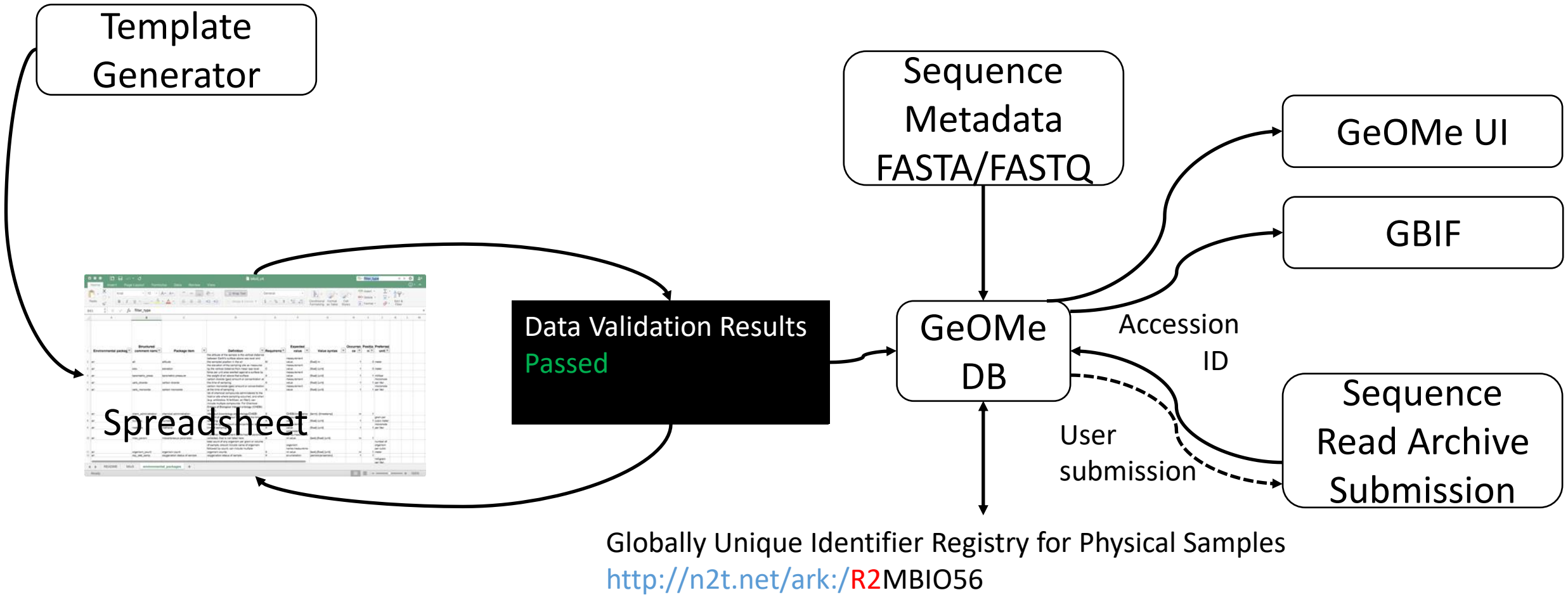
Support for non-human variant data archival and accessioning is transitioning from dbSNP to EVA from September 2017.

Data requirements

EVA accepts submission of genetic variation data based on three criteria:

1. The genome assembly used is International Nucleotide Sequence Database Collaboration (INSDC) registered, or will be at point of submission
2. The variation data is described in valid VCF file(s) this can be tested prior to submission using the EVA VCF Validation Suite found here
3. For all data submitted to the EVA, we require that it be possible to compute allele frequencies for all submitted variants. Therefore, the EVA supports two types of submissions: 1) variation data with sample genotypes 2) summary data with population allele frequencies

Genomic Observatories MetaDatabase Workflow



GeOMe adopts terms from standards + a minimum set of required fields.

Biodiversity Information Standards TDWG

Darwin Core

MIxS

gsc genomic STANDARDS consortium

NCBI

Envo

GeOMe Projects adopts specific terms and custom validation rules for specific use cases.

e.g. requiring trait descriptions, or environmental measurements.

ARMS

GLOBAL ARMS PROGRAM

Diversity of the Indo-Pacific Network