Linking genomic to specimen & species trait data

Episode 1: a winge...

Episode 2: Something cool...

With thanks to Paula Mabee!
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
TARGET CAPTURE SEQUENCE DATA: the bioinformatics grey zone

SNPs eg ddRAD, GbS

Target capture

UCEs

Anchored Enrichment

Exons

Custom exons

RNASeq

Whole Genome Resequecing

10K avian species, 1000s of UCE loci

700 species, >2000 specimens, ~2K exons
Exon capture to link micro to macrtoevolution: Eugongylus skinks (Bragg et al. 2018, Potter et al. 2018)

~200 taxa, 1K specimens, ~3K exons
Phylogenomics of a rapid radiation: the Australian rainbow skinks

Jason G. Bragg, Sally Potter, Ana C. Afonso Silva, Conrad J. Hoskin, Benjamin Y. H. Bai and Craig Moritz

ncbi SRA Bioproject

Raw data

Code & alignments

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


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

https://www.geome-db.org/
Correcting the disconnect between phylogenetics and biodiversity informatics

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

phylojive.ala.org.au

Only for tips = accepted names
Example:
Morphological evolution in myobatrachid frogs

Vidal-Garcia et al. 2014
J. Evol. Biol. 27:181-192
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.
Support for non-human variant data archival and accessioning is transitioning from dbSNP to EVA from September 2017.
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
Example: *H. binoei* lineages

[Live URL for *H. binoei* example]
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

Template Generator

Spreadsheet

Data Validation Results Passed

Sequence Metadata FASTA/FASTQ

GeOMe UI

GBIF

Sequence Read Archive Submission

Accession ID

User submission

Globally Unique Identifier Registry for Physical Samples
http://n2t.net/ark:/R2MBIO56
**GeOMe** adopts terms from standards + a minimum set of required fields.

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

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