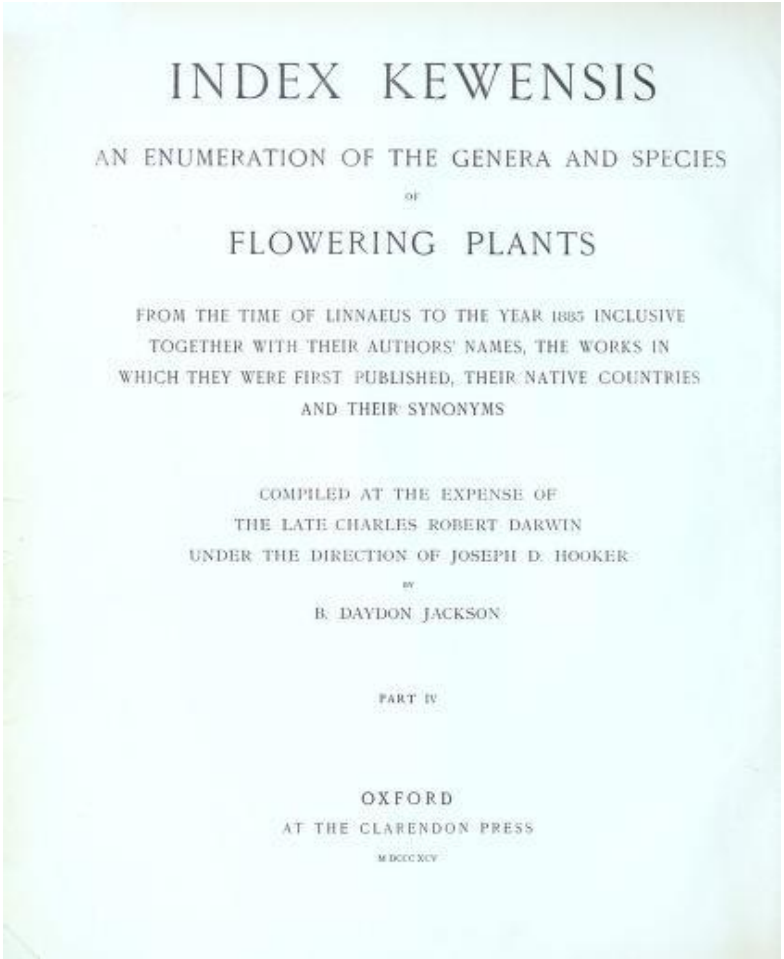
Natural History Collections in the Digital Age

Paul Kersey
Deputy Director of Science (Research)
Royal Botanic Gardens, Kew

- 250 years of science at Kew
  - Scientific mission focused on plant and fungal diversity



INDEX KEWENSIS

AN ENUMERATION OF THE GENERA AND SPECIES
OF
FLOWERING PLANTS

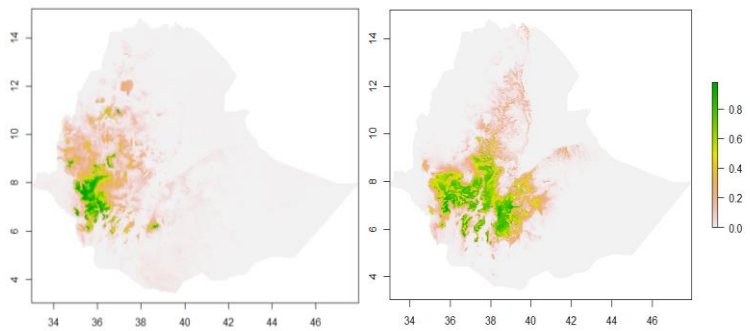
FROM THE TIME OF LINNAEUS TO THE YEAR 1885 INCLUSIVE
TOGETHER WITH THEIR AUTHORS' NAMES, THE WORKS IN
WHICH THEY WERE FIRST PUBLISHED, THEIR NATIVE COUNTRIES
AND THEIR SYNONYMS

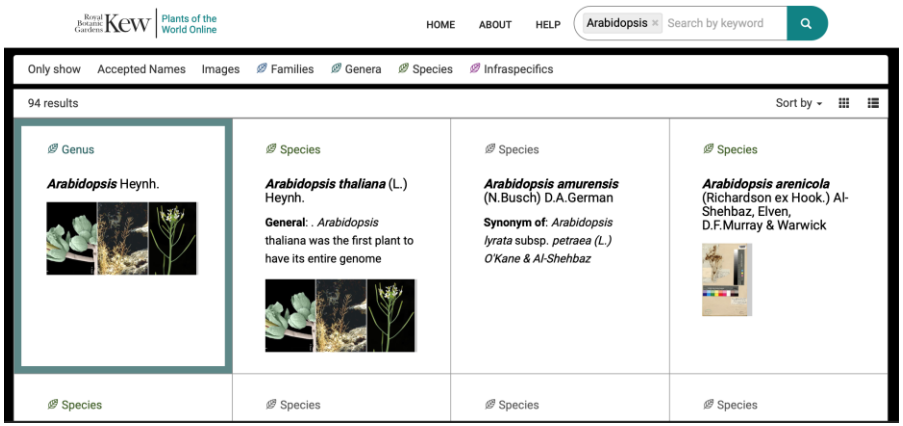
COMPILED AT THE EXPENSE OF
THE LATE CHARLES ROBERT DARWIN
UNDER THE DIRECTION OF JOSEPH D. HOOKER
BY
B. DAYDON JACKSON

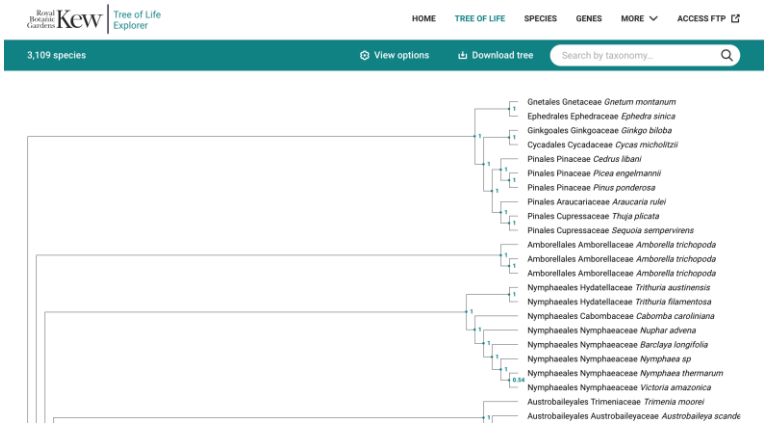
PART IV

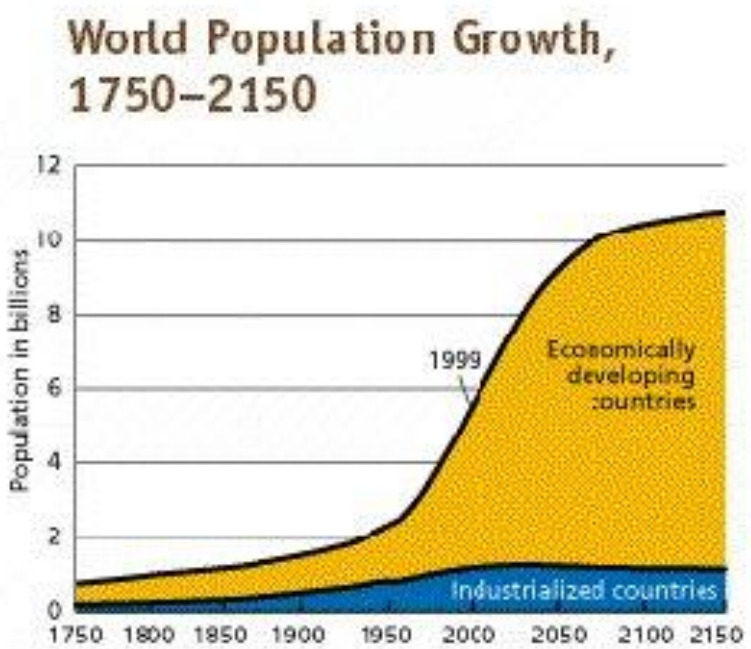
OXFORD
AT THE CLARENDON PRESS
M DCCC XCV

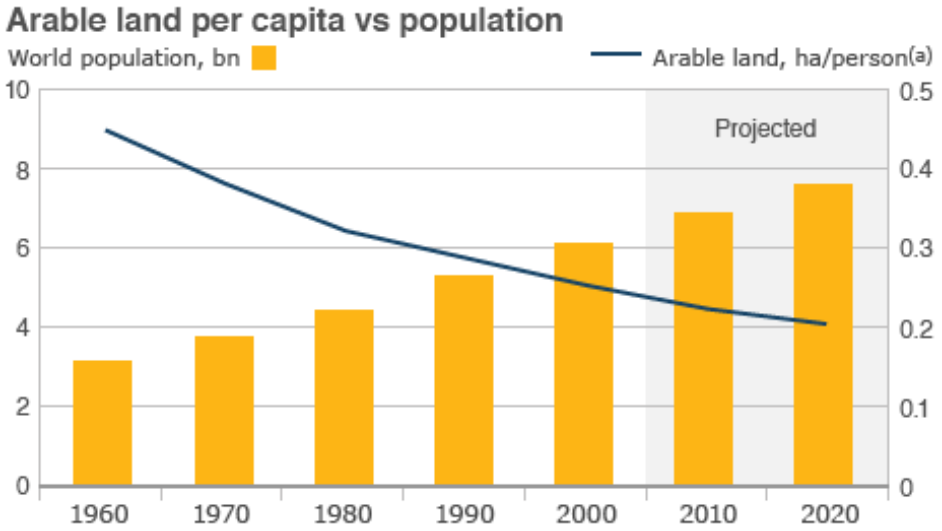
Natural History Collections in the Digital Age











(a) Hectares per person of arable land and land in permanent crops
Source: BHP Billiton investor presentation

Weatherwatch: melting Arctic ice brings US heatwaves

Study finds warm springtimes in Hudson Bay lead to hot summers further south

▲ Sea ice in the Hudson Strait. Arctic ice has shrunk drastically in the last 40 years. Photograph: Kike Calvo/National Geographic Creative/Alamy Stock Photo

Arctic sea ice has diminished drastically over the last 40 years. Now a study shows that the southern half of the US may be suffering from more summer heatwaves as a consequence.

Dagmar Budikova, a climatologist at Illinois State University, and colleagues used satellite data to measure the ebb and flow of Arctic sea ice in the

‘Frightening’ number of plant extinctions found in global survey

Study shows 571 species wiped out, and scientists say figure is likely to be big underestimate

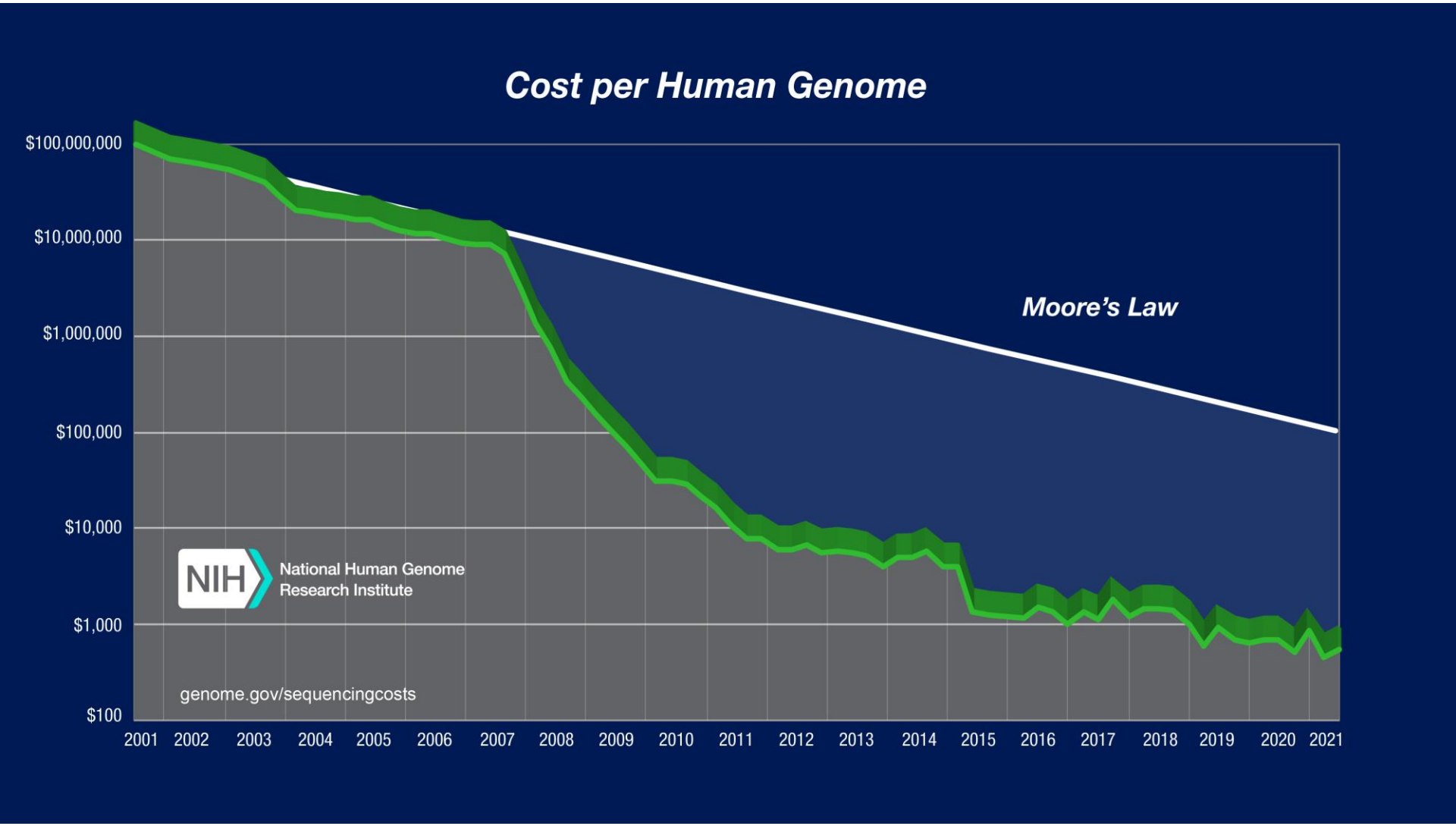
▲ Maria Vorontsova: ‘We take them for granted and I don’t think we should.’ Photograph: Kew Gardens

Human destruction of the living world is causing a “frightening” number of plant extinctions, according to scientists who have completed the first global analysis of the issue.

They found 571 species had definitely been wiped out since 1750 but with knowledge of many plant species still very limited the true number is likely to be much higher. The researchers said the plant extinction rate was 500

- The future of herbaria, fungaria, and living/seed collections
  - Preserve living biodiversity *ex situ*
    - For re-introduction
    - As a source of material for research and exploitation
  - Preserved collections as bio-archives
    - A measurement means nothing without material
    - Irreplaceable historical record of what occurred where/when

- To be useful, the collections must be digitized
  - Catalogue, image what we have
  - Web portals and programmatic interfaces to provide global access
  - Molecular digitization (i.e., genomic and chemical characterization)
  - Machine learning


Cost per Human Genome
$100,000,000
$10,000,000
$1,000,000
$100,000
$10,000
$1,000
$100
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021

- Genomics can tell us:
  - What something is (definition, identification)
  - How, and why, it is different to, and similar to, other things (taxonomy, evolution, function)
  - The composition of communities, habitats, and population and how these are changing over time
    - And what each species is doing in each ecosystem

Phylogenetics of Seed Plants: An Analysis of Nucleotide Sequences from the Plastid Gene rbcL

Author(s): Mark W. Chase, Douglas E. Soltis, Richard G. Olmstead, David Morgan, Donald H. Les, Brent D. Mishler, Melvin R. Duvall, Robert A. Price, Harold G. Hills, Yin-Long Qiu, Kathleen A. Kron, Jeffrey H. Rettig, Elena Conti, Jeffrey D. Palmer, James R. Manhart, Kenneth J. Sytsma, Helen J. Michaels, W. John Kress, Kenneth G. Karol, W. Dennis Clark, Mikael Hedren, Brandon S. Gaut, Robert K. Jansen, Ki-Joong Kim, C... James F. Smith...

Source: *Annals of the Missouri Botanical Garden*, Vol. 80, No. 3 (1993

Published by: Missouri Botanical Garden Press

Stable URL: https://www.jstor.org/stable/2399846

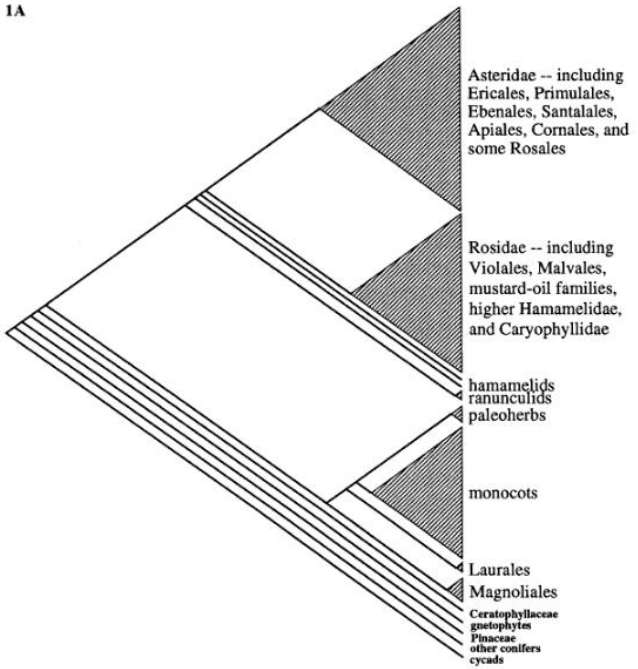
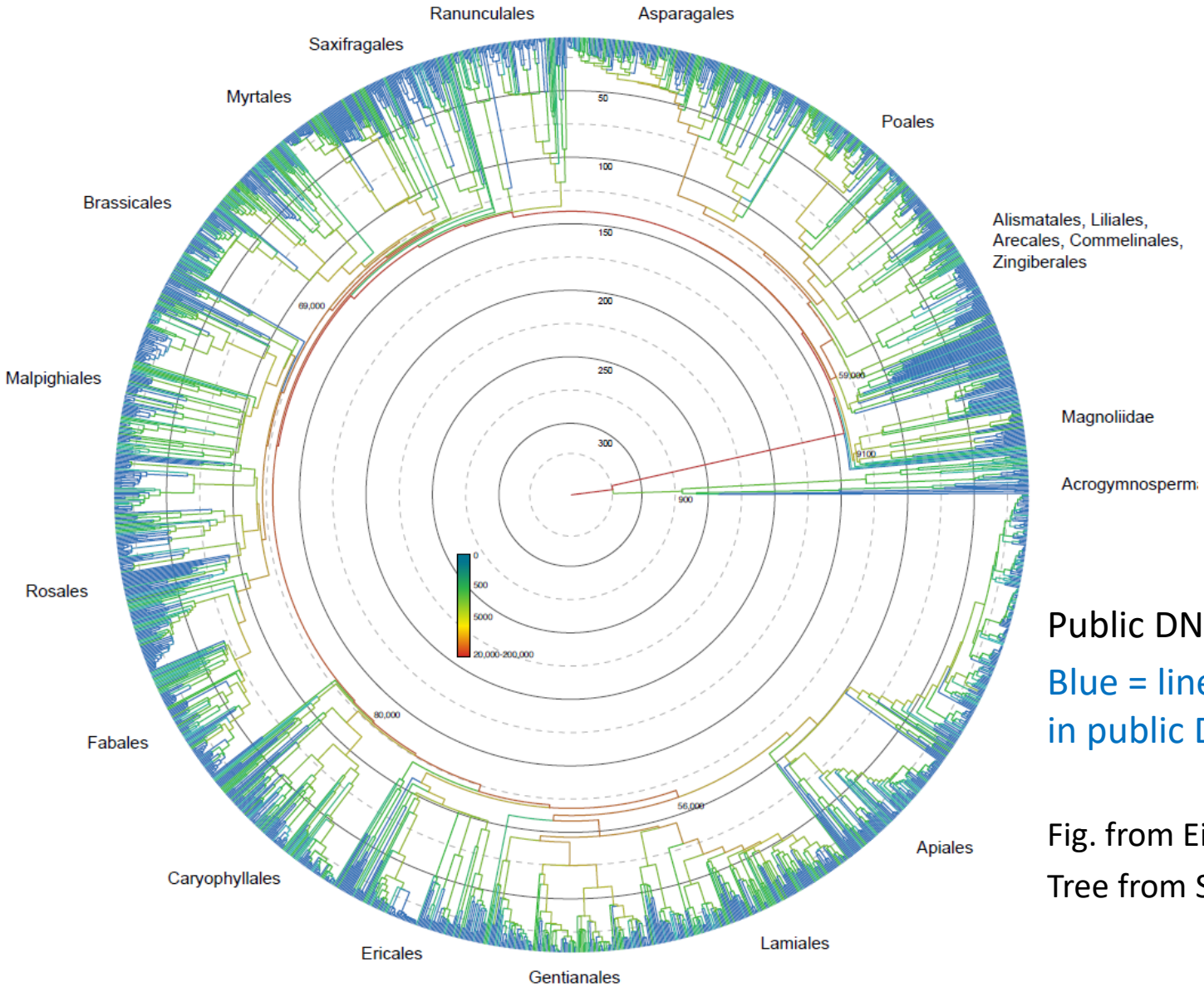


FIGURE 1. Summaries of the major clades identified in: (A) the combinable component consensus tree of 500 equally parsimonious trees found for 475 taxa using the character-state weighting method of Albert et al. (1993, this issue); and (B) the strict consensus tree of 3,900 equally parsimonious trees for 499 taxa found using the Fitch (even weights) criterion. These are ingroup networks arranged arbitrarily with the cycads sister to all other seed plants.



Public DNA data for seed plants

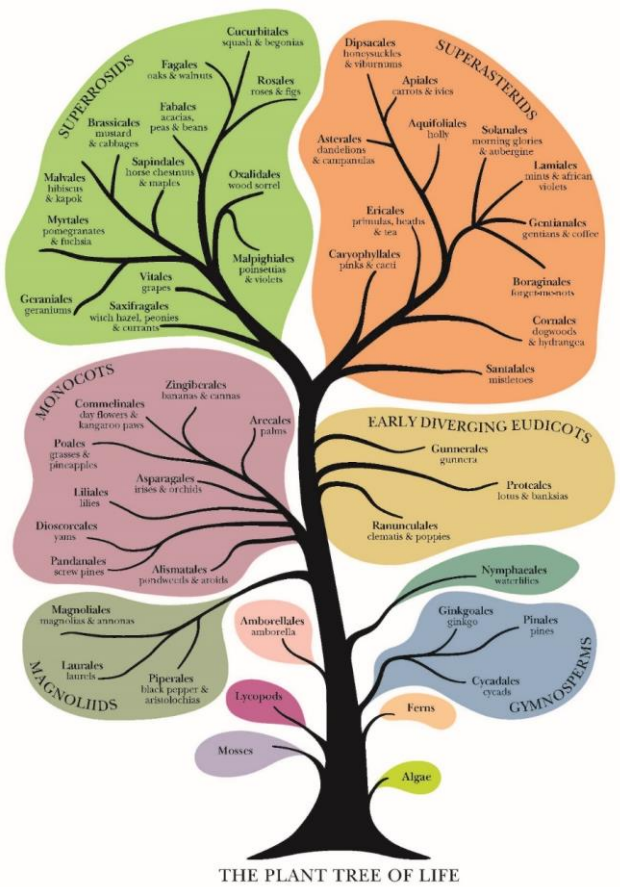
Blue = lineage not represented in public DNA database

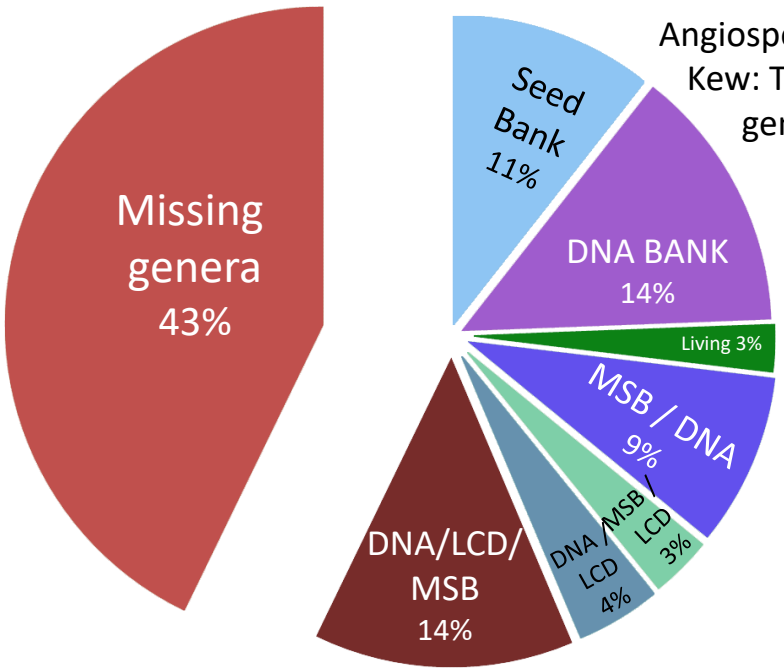
Fig. from Eiserhardt et al. 2018 AJB

Tree from Smith & Brown 2018 AJB

Plant and Fungal Trees of Life (PAFTOL)

- Genomic characterisation of Kew's collections
- Genome-scale data from each genus (*ca* 24,000 genera, including *ca* 24,000 plants, 8,200 fungi)
- Complete genus-level phylogenies for plants & fungi
- Dynamically updated via a web interface





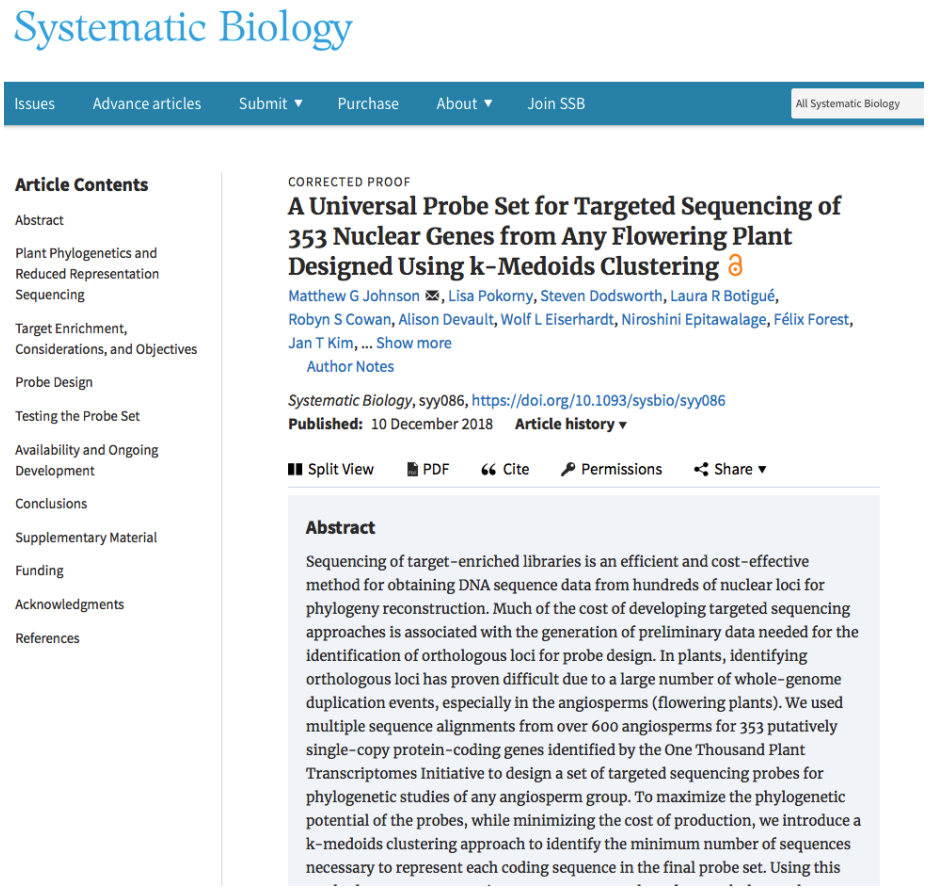
Angiosperm samples at Kew: Total available genera: 57%

Magic mushroom (*Psilocybe cyanescens*)
The fungal tree of life is the key to understanding the properties and dangers of this iconic species

Angiosperms353 Probe Set

- Breadth – one kit for all angiosperms
- Depth – works from higher to species level (even barcoding?)
- Effective with degraded DNA

- 353 low-copy loci (260kbp)
- Designed from 1KP transcriptome data (600 genera, 209 families, 80000 probes)
- 5-15 sequences selected per locus
- Median recovery: 137kbp target loci (max 250kbp) plus 212kbp intron
- Available now at **one third of the price** of a custom bait kit
- www.arborbiosci.com/products/mybaits-plant-angiosperms/



Systematic Biology

Issues | Advance articles | Submit | Purchase | About | Join SSB | All Systematic Biology

Article Contents

Abstract
Plant Phylogenetics and Reduced Representation Sequencing
Target Enrichment, Considerations, and Objectives
Probe Design
Testing the Probe Set
Availability and Ongoing Development
Conclusions
Supplementary Material
Funding
Acknowledgments
References

CORRECTED PROOF

A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using k-Medoids Clustering

Matthew G Johnson, Lisa Pokorny, Steven Dodsworth, Laura R Botigué, Robyn S Cowan, Alison Devault, Wolf L Eiserhardt, Niroshini Epitawalage, Félix Forest, Jan T Kim, ... Show more
Author Notes

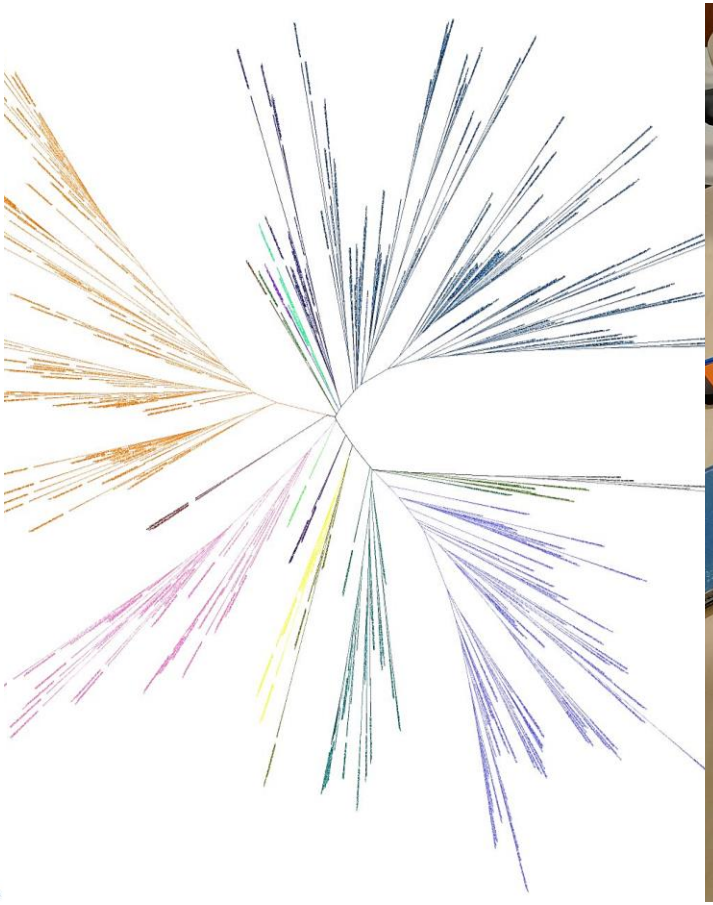
Systematic Biology, syy086, https://doi.org/10.1093/sysbio/syy086
Published: 10 December 2018 **Article history**

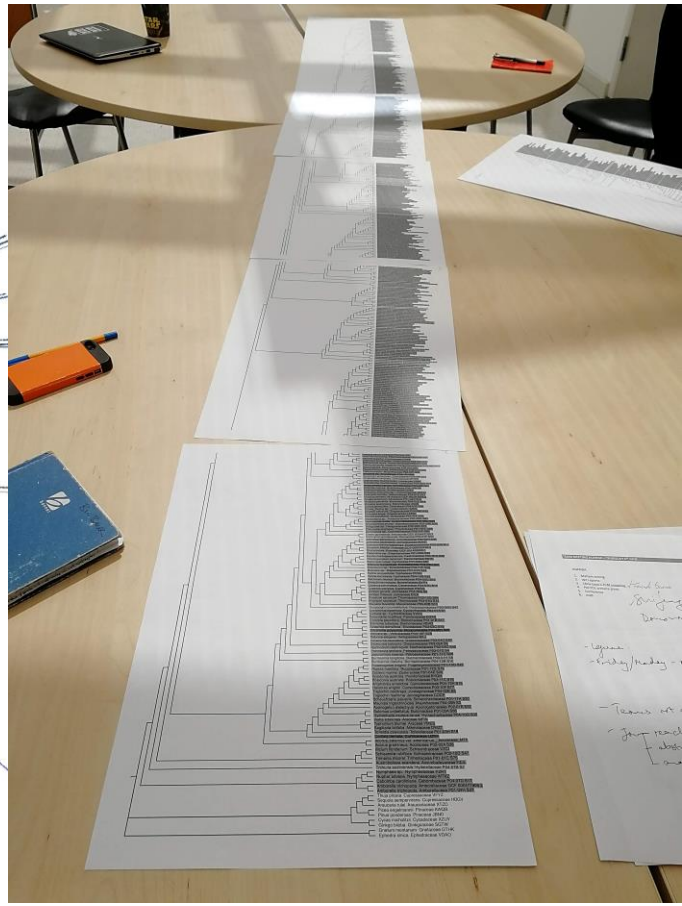
Split View | PDF | Cite | Permissions | Share

Abstract

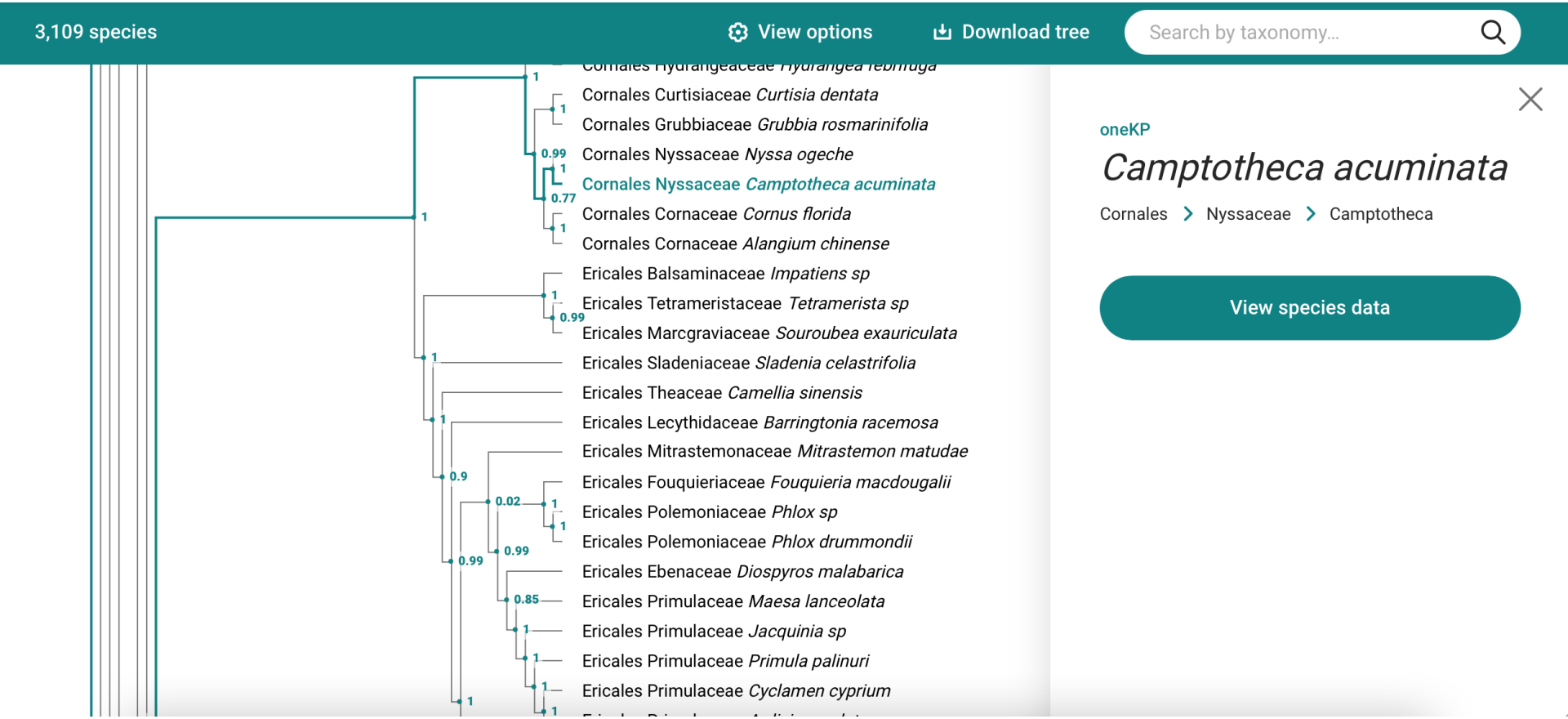
Sequencing of target-enriched libraries is an efficient and cost-effective method for obtaining DNA sequence data from hundreds of nuclear loci for phylogeny reconstruction. Much of the cost of developing targeted sequencing approaches is associated with the generation of preliminary data needed for the identification of orthologous loci for probe design. In plants, identifying orthologous loci has proven difficult due to a large number of whole-genome duplication events, especially in the angiosperms (flowering plants). We used multiple sequence alignments from over 600 angiosperms for 353 putatively single-copy protein-coding genes identified by the One Thousand Plant Transcriptomes Initiative to design a set of targeted sequencing probes for phylogenetic studies of any angiosperm group. To maximize the phylogenetic potential of the probes, while minimizing the cost of production, we introduce a k-medoids clustering approach to identify the minimum number of sequences necessary to represent each coding sequence in the final probe set. Using this


Gnetidae
Cycadidae
Ginkgoidae
Pinidae
Amborellales
Nymphaeales
Austrobaileyales
Ceratophyllales
Acorales
Alismatales
Burmanniaceae
Petrosaviales
Dioscoreales
Pandanales
Poales
Dasypogonaceae
Arecales
Pontederiaceae
Commelinales
Philydraceae
Haemodoraceae
Hanguanaceae
Zingiberales
Liliales
Asparagales
Piperales
Canellales
Laurales
Magnoliales
Chloranthales
Proteales
Ranunculales
Buxales
Trochodendrales
Nelumbonaceae
Caryophyllales
Berberidopsidales
Cornales
Ericales
Cardiopteridaceae
Metteniusales
Forgesia racemosa
Brunialis
Escalloniales
Paracryphiales
Pennantiaceae
Dipsacales
Apiales
Asterales
Oncothecaceae
Garryales
Aquifoliales
Icacinales
Hydroleaceae
Vahliales
Montiniaceae
Boraginales
Gentianales
Solanales
Lamiales
Dilleniales
Gunnerales
Vitales
Santalales
Saxifragales
Myrtales
Zygophyllales
Geraniales
Crossosomatales
Picramniales
Cucurbitales
Fabales
Fagales
Rosales
Celastrales
Malpighiales
Huerteales
Oxalidales
Sapindales
Malvales
Brassicales





- http://treeoflife.kew.org provides access to angiosperm data
- Release 2.0 (January 2022) contains 9834 samples, 7514 genera


3,109 species
View options
Download tree
Search by taxonomy...
Cornales Curtisiaceae Curtisia dentata
Cornales Grubbiaceae Grubbia rosmarinifolia
Cornales Nyssaceae Nyssa ogeche
Cornales Nyssaceae Camptotheca acuminata
Cornales Cornaceae Cornus florida
Cornales Cornaceae Alangium chinense
Ericales Balsaminaceae Impatiens sp
Ericales Tetrameristaceae Tetramerista sp
Ericales Marcgraviaceae Souroubea exauriculata
Ericales Sladeniaceae Sladenia celastrifolia
Ericales Theaceae Camellia sinensis
Ericales Lecythidaceae Barringtonia racemosa
Ericales Mitrastemonaceae Mitrastemon matudae
Ericales Fouquieriaceae Fouquieria macdougalii
Ericales Polemoniaceae Phlox sp
Ericales Polemoniaceae Phlox drummondii
Ericales Ebenaceae Diospyros malabarica
Ericales Primulaceae Maesa lanceolata
Ericales Primulaceae Jacquinia sp
Ericales Primulaceae Primula palinuri
Ericales Primulaceae Cyclamen cyprium
oneKP
Camptotheca acuminata
Cornales › Nyssaceae › Camptotheca
View species data

Showing 20 of 3,109 species | Export table | Search by taxonomy...

oneKP	Gnetales	Gnetaceae	Gnetum	*Gnetum montanum*	30
PAFTOL	Lamiales	Acanthaceae	Afrofittonia	*Afrofittonia silvestris*	344
oneKP	Lamiales	Acanthaceae	Anisacanthus	*Anisacanthus quadrifidus*	343
PAFTOL	Lamiales	Acanthaceae	Anisosepalum	*Anisosepalum alboviolaceum*	337
PAFTOL	Lamiales	Acanthaceae	Anisostachya	*Anisostachya tenella*	344
PAFTOL	Lamiales	Acanthaceae	Aphelandra	*Aphelandra fulgens*	339
PAFTOL	Lamiales	Acanthaceae	Asystasia	*Asystasia leptostachya*	343
PAFTOL	Lamiales	Acanthaceae	Avicennia	*Avicennia marina*	345
PAFTOL	Lamiales	Acanthaceae	Barleria	*Barleria ventricosa*	328
PAFTOL	Lamiales	Acanthaceae	Blepharis	*Blepharis sp.*	346

PAFTOL

Anisostachya tenella

Lamiales > Acanthaceae > Anisostachya

Molmou, D. 1021 (K) | K001243667

Download gene sequences

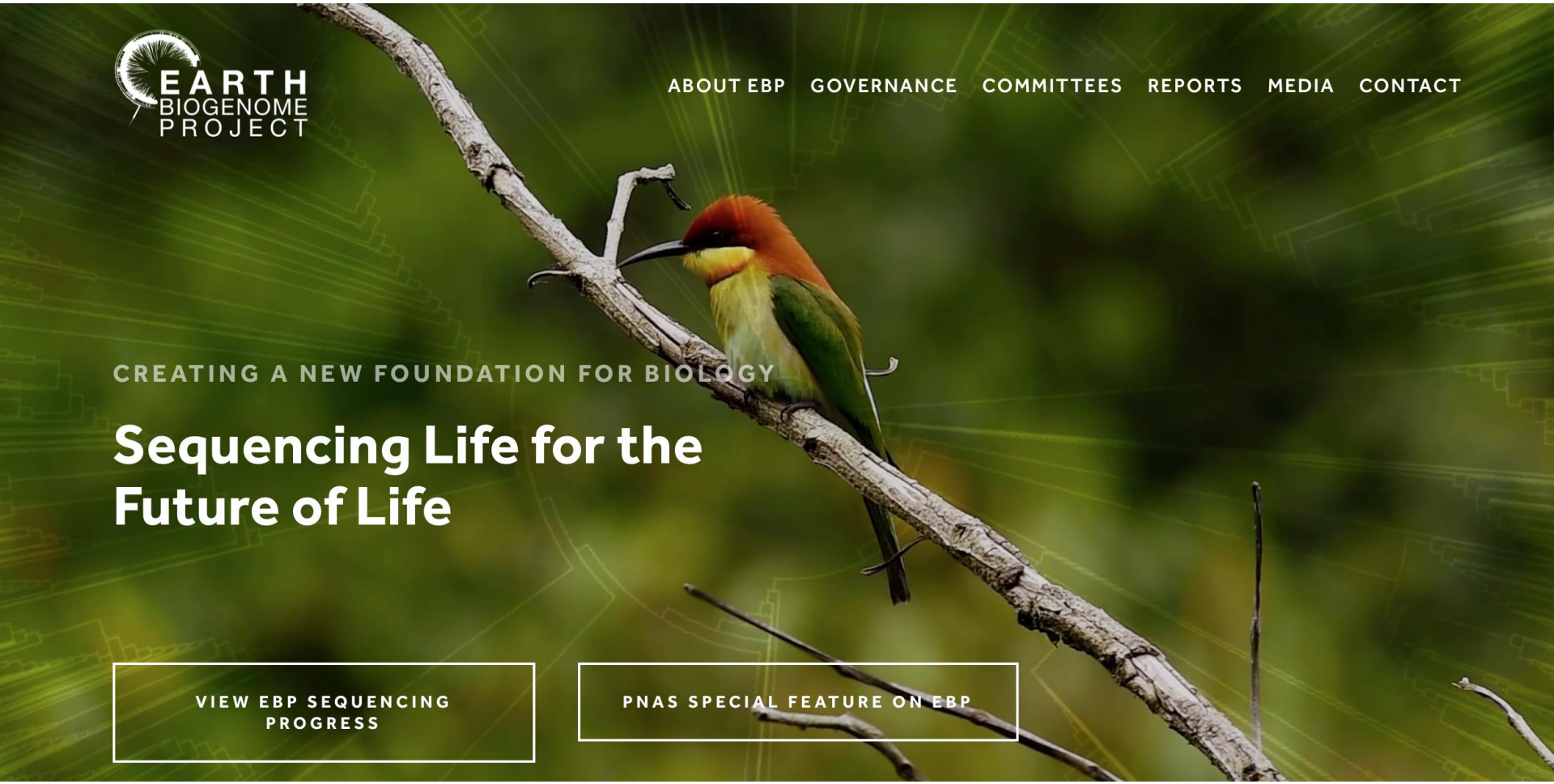
View in tree of life

Primary data

No. of reads: 8,621,034

Sequencing platform: HiSeq

Data access: ERX4841148


EARTH BIOGENOME PROJECT
ABOUT EBP
GOVERNANCE
COMMITTEES
REPORTS
MEDIA
CONTACT
CREATING A NEW FOUNDATION FOR BIOLOGY
Sequencing Life for the Future of Life
VIEW EBP SEQUENCING PROGRESS
PNAS SPECIAL FEATURE ON EBP

THE EBP ROADMAP CALLS FOR SEQUENCING AND ANNOTATING ~1.5 MILLION KNOWN EUKARYOTIC SPECIES IN THREE PHASES OVER A 10-YEAR PERIOD USING A PHYLOGENOMIC APPROACH.

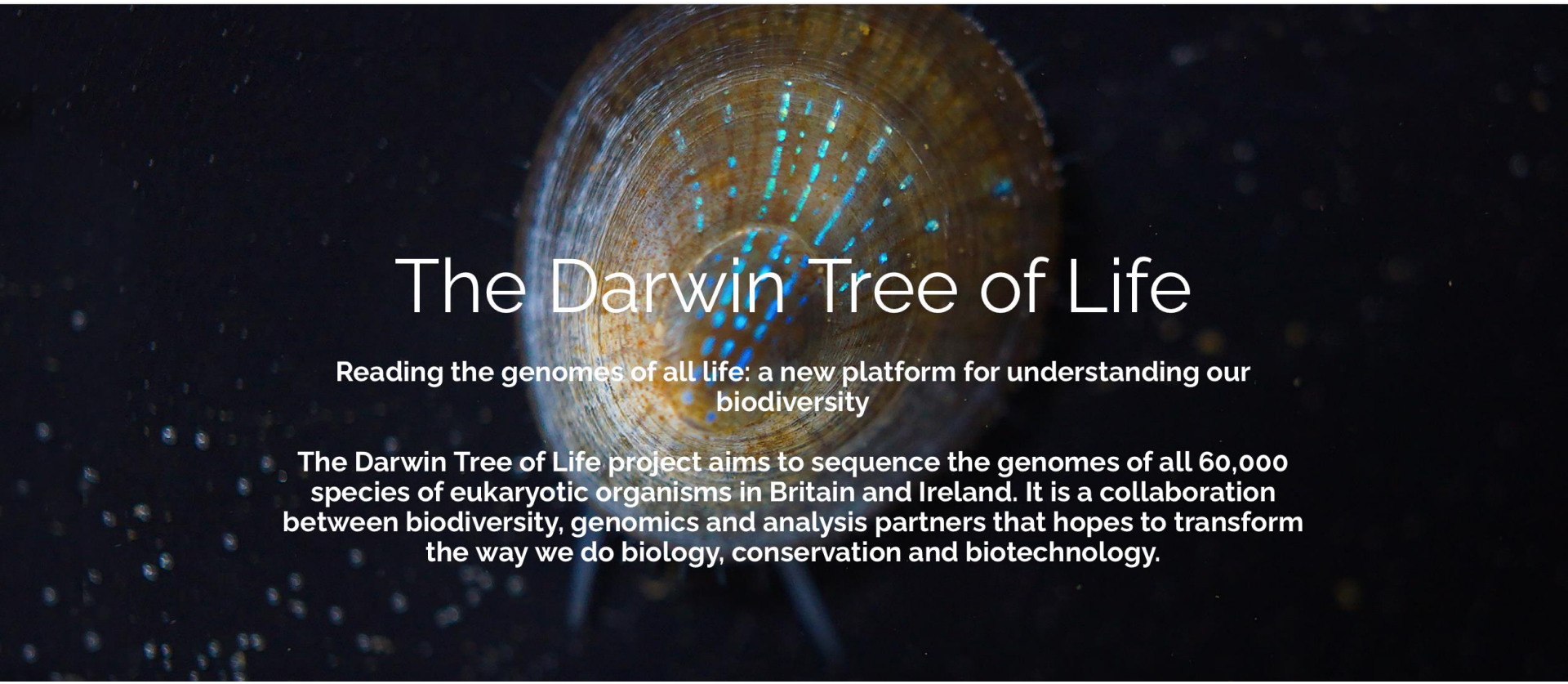
During the three years of Phase I, one of the most important goals is to create annotated chromosome-scale reference assemblies for at least one representative species of each of the ~9,000 eukaryotic taxonomic families. Nucleotide divergence and divergence time will be additional factors in the selection of species so that balance across eukaryotic taxa is achieved. High-quality reference assemblies at the family level will ensure robustness of comparative genomic analyses by providing complete gene sets as well as ordered and oriented syntenic blocks created by genome scaffolding methods. In addition, these genomes will be useful for classification of extant and new species, identification of genetic changes associated with specialized traits in specific lineages, in silico reference-assisted scaffolding of assemblies produced in Phase II and Phase III of the project, in silico reconstruction of ancestral genomes, and rescue of species from extinction. A full description of the roadmap, overall strategy, and estimated costs can be found in the *PNAS Perspective* SI Appendix.

- Earth Biogenome project: rationale
  - Exploit biological dark matter for human benefit
    - Find useful molecules (food, fuel, biocatalysts, pharmaceuticals, etc.)
  - Protect ecosystems by characterising them, conserving their biological diversity, understanding them, and reducing external pressures upon them
  - Accelerate scientific understanding by placing each focus of study within its full evolutionary context
    - Genus-level sequencing is not enough: *Astragalus* has more than 3,200 species





- Earth Biogenome project
  - 10-15 million eukaryotic species
  - ~350,000 vascular plants
  - Only 2-3 million described
  - Only 2,500 sequenced
  - Plan: sequence 1.5 million species over a 10-year horizon


Darwin
TREE of LIFE
HOME
DATA
JOBS
BLOG
PROJECT RESOURCES
The Darwin Tree of Life
Reading the genomes of all life: a new platform for understanding our biodiversity
The Darwin Tree of Life project aims to sequence the genomes of all 60,000 species of eukaryotic organisms in Britain and Ireland. It is a collaboration between biodiversity, genomics and analysis partners that hopes to transform the way we do biology, conservation and biotechnology.

- Darwin Tree of Life: Pilot Phase
  - Aim: sequence ~60K British species within 7-10 years
  - Pump priming phase
    - Aim to bank 8K species, and sequence 6K of these
  - Aim for G10K/Earth Biogenome assembly standards (3.4.2.Qv40phased metric)
    - *Minimum contig N50 of 1 million bp (1Mb)*
    - *Scaffold N50 of 10Mb*
    - *90% of the genome assembled into chromosomes confirmed by 2 independent sources*
    - *Base-call quality error of QV40 (no more than 1 nucleotide error in 10,000 bp)*
    - *Haplotype phased*.

Darwin Tree of Life: Pilot Phase Targets

Taxon	Species count	2019		2020		2021		Total		average size Gb	total Gb	2022+
Deuterostomia	1625	145	25	400	82	553	213	1098	320	1.6	527	1299
Arthropoda	26520	544	110	1450	300	2300	540	4294	950	0.8	732	25564
Other Metazoa	7200	90	10	300	75	750	150	1140	235	1.2	285	6965
Plants	3226	120	20	345	100	350	170	815	290	1.6	473	2936
Fungi	16100	100	15	220	30	520	70	840	115	0.3	35	15985
"Protista"	5000	20	10	80	40	100	40	200	90	0.5	45	4910
Totals	59677	1019	190	2795	627	4573	1183	8387	2000	1.0	2096	57659

Current status: 45 genomes assembled, 3 published

DATA NOTE

The genome sequence of the chicken of the woods fungus, *Laetiporus sulphureus* (Bull.) Murrill, 1920 [version 1; peer review: awaiting peer review]

Richard Wright [1,2], Kieran Woof [1], Brian Douglas[1], Ester Gaya [1],
Royal Botanic Gardens Kew Genome Acquisition Lab, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific

Operations: DNA Pipelines collective, Tree of Life Core Informatics collective, Darwin Tree of Life Consortium

Author details

Abstract

We present a genome assembly from an individual *Laetiporus sulphureus* (the chicken of the woods fungus; Basidiomycota; Agaricomycetes; Polyporales; Laetiporaceae). The genome sequence is 37.4 megabases in span. The complete assembly is scaffolded into 14 chromosomal pseudomolecules.

Keywords

Digitisation of the Herbarium & Fungarium Specimens –

- **To digitise Kew's 8.5 million plant and fungal specimens and create a complete catalogue of our internationally significant collections**, making specimen records and images freely available online and accessible to researchers across the globe.
- The result will be a unique, world-leading resource. It will open up data from more than 260 years of scientific exploration, placing them at the centre of efforts to combat urgent global challenges such as habitat degradation, climate change and human health

Integrated Collection Management System (ICMS)

- Ensure efficient and effective integrated management of the Science and Living Collections and the data describing them. This will allow more efficient tracking of the use of specimens to meet legislative requirements concerning access to genetic resources and benefit sharing, including the Nagoya Protocol of the Convention on Biological Diversity and plant health legislation.

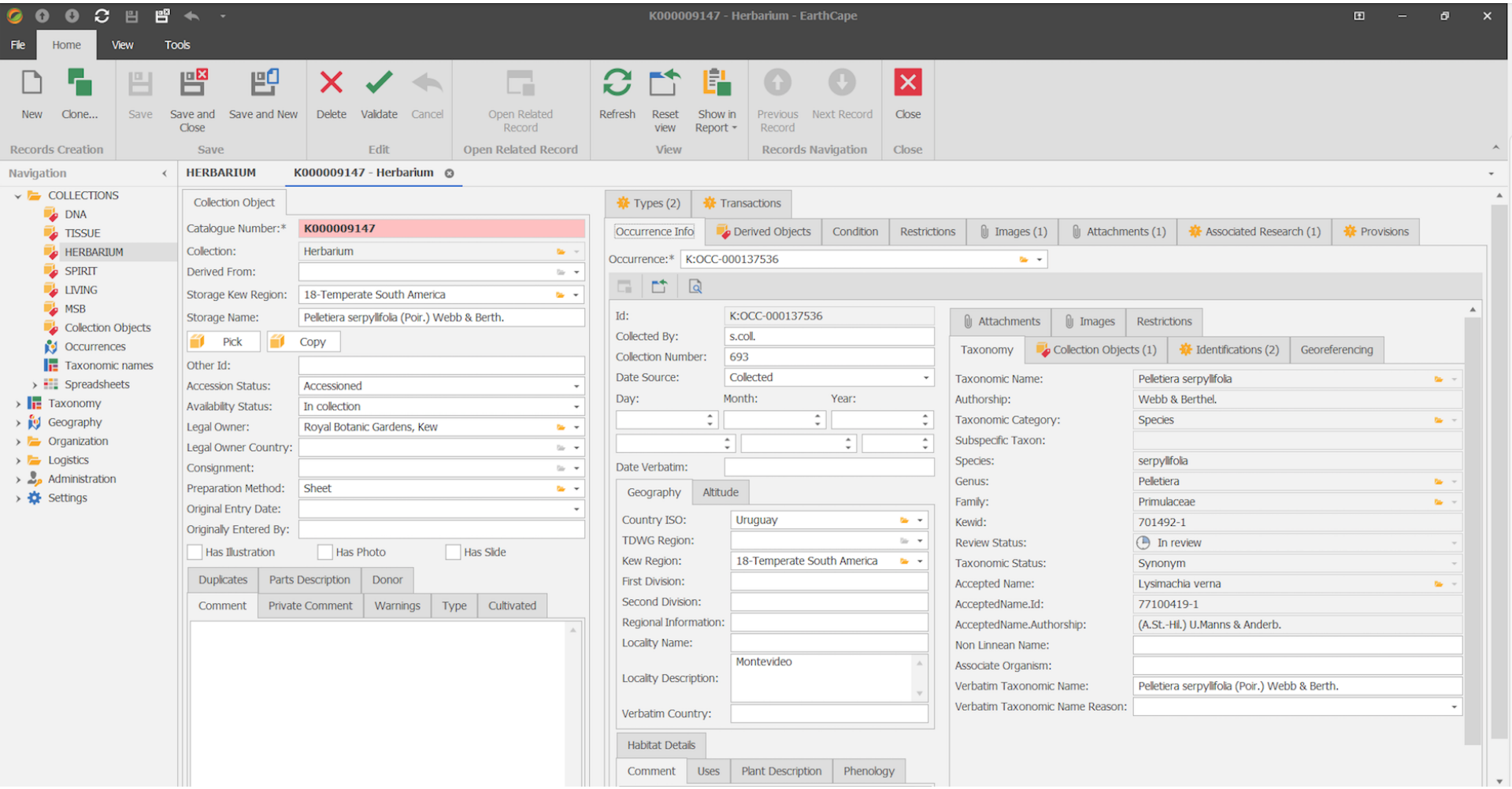
Data Portal

- The data portal will provide the web-based interface to allow open access to our digitised Collection data

HERBARIUM	Y1	Y2	Total Y1-Y2	Y3	Y4	Total Y3-Y4	Total
Specimen Images	1,160,000	2,084,000	3,244,000	2,084,000	1,172,000	3,256,000	6,500,000
Folder Images	406,000	729,400	1,135,400	729,400	410,200	1,139,600	2,275,000
Folder Transcriptions	830,000	1,320,000	2,150,000	2,175,000	2,175,000	4,350,000	6,500,000
Specimen Label Transcriptions	700,000	1,320,000	2,020,000	1,990,000	1,990,000	3,980,000	6,000,000
Complete Digital Specimen Record	**700,000**	**1,320,000**	**2,020,000**	**1,990,000**	**1,990,000**	**3,980,000**	**6,000,000**

FUNGARIUM	Y1	Y2	Total Y1&2	Y3	Y4	Total Y3&4	Total
Specimen Label Transcriptions	0	200,000	200,000	250,000	275,000	525,000	725,000
Complete Digital Specimen Record	**0**	**200,000**	**200,000**	**250,000**	**275,000**	**525,000**	**725,000**

Royal Botanic Gardens Kew



Home > Occurrence records > Record: Herbarium:K000318424 (Iberis odorata L.)

Occurrence record: Herbarium:K000318424

Specimen of *Iberis odorata L.* L. recorded on 1975-04-19

Flag an issue | Copy record id | Compare "original vs processed" values

- Dataset
- Event
- Taxonomy
- Geospatial
- Images
- Data quality tests (1, 3, 24, 10, 48)

Images

Rights holder: Royal Botanic Gardens, Kew
Image license: http://creativecommons.org/licenses/by/4.0/
View image details

Date loaded: 2022-01-14
Date last processed: 2022-01-14

Dataset

Data resource	Royal Botanic Gardens, Kew - Herbarium Specimens
Institution code	Royal Botanic Gardens, Kew *Supplied institution code "K"*
Collection code	Herbarium *Supplied collection code "Herbarium"*
Catalog number	K000318424
Occurrence ID	http://specimens.kew.org/herbarium/K000318424
Basis of record	Preserved specimen
Identified by	Sahra
Collector	Staff, B. Botany *Supplied as "Botany staff"*
Record number	43397
Occurrence status	present
Abcd identification qualifier	Not provided

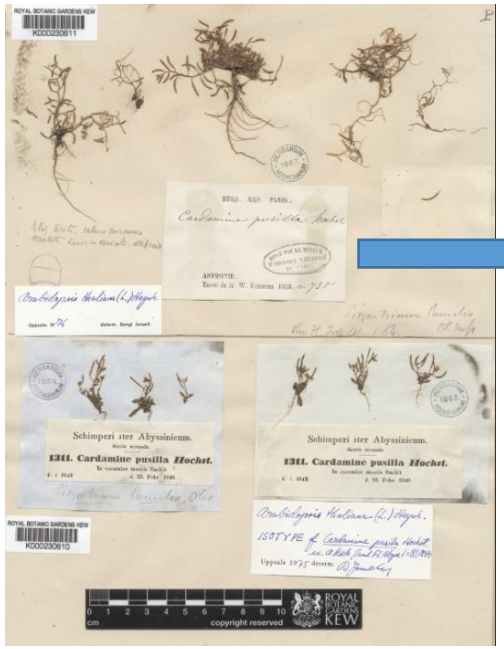
Event

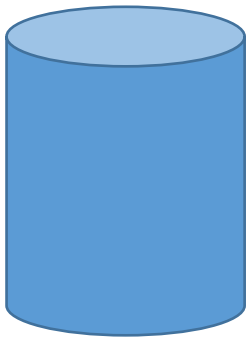
Occurrence date	1975-04-19
Date precision	Day

Taxonomy

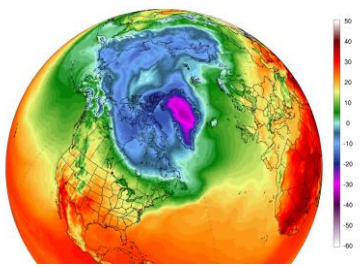
Higher classification	CRUCIFERAE
Scientific name	Iberis odorata L.
Taxon rank	[rank not known]
Family	CRUCIFERAE
Genus	*Iberis*
Species	*Iberis odorata*
Taxonomic issue	No issues
Name match metric	No match
Name parse type	SCIENTIFIC
Scientific name authorship	L.

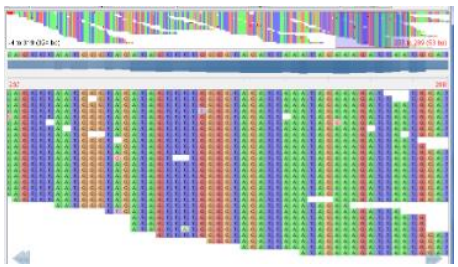
The Digitized Herbarium/Fungarium

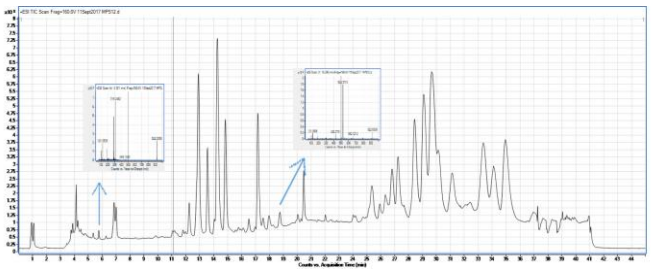




Notes: *Manjekia* is a moderately robust, solitary palm that is recorded from several lowland rainforest localities on limestone in Biak Island. Its arching leaves bear broadly lanceolate, pendulous, praemorse leaflets with concave, praemorse apices and its inflorescence axes are white or greenish white. *Manjekia* bears some similarities to *Adonidia*, in which it was originally described, such as the white inflorescences branched up to four orders, the red fruit and similar seed and endocarp morphology, but its leaves do not resemble those of *Adonidia*, which have ascending, narrow leaflets in slightly different planes that are less conspicuously praemorse and fewer stamens in the staminate flowers (30–32, instead of 45–50 in *Adonidia*). See also notes under *Jailoloa*.

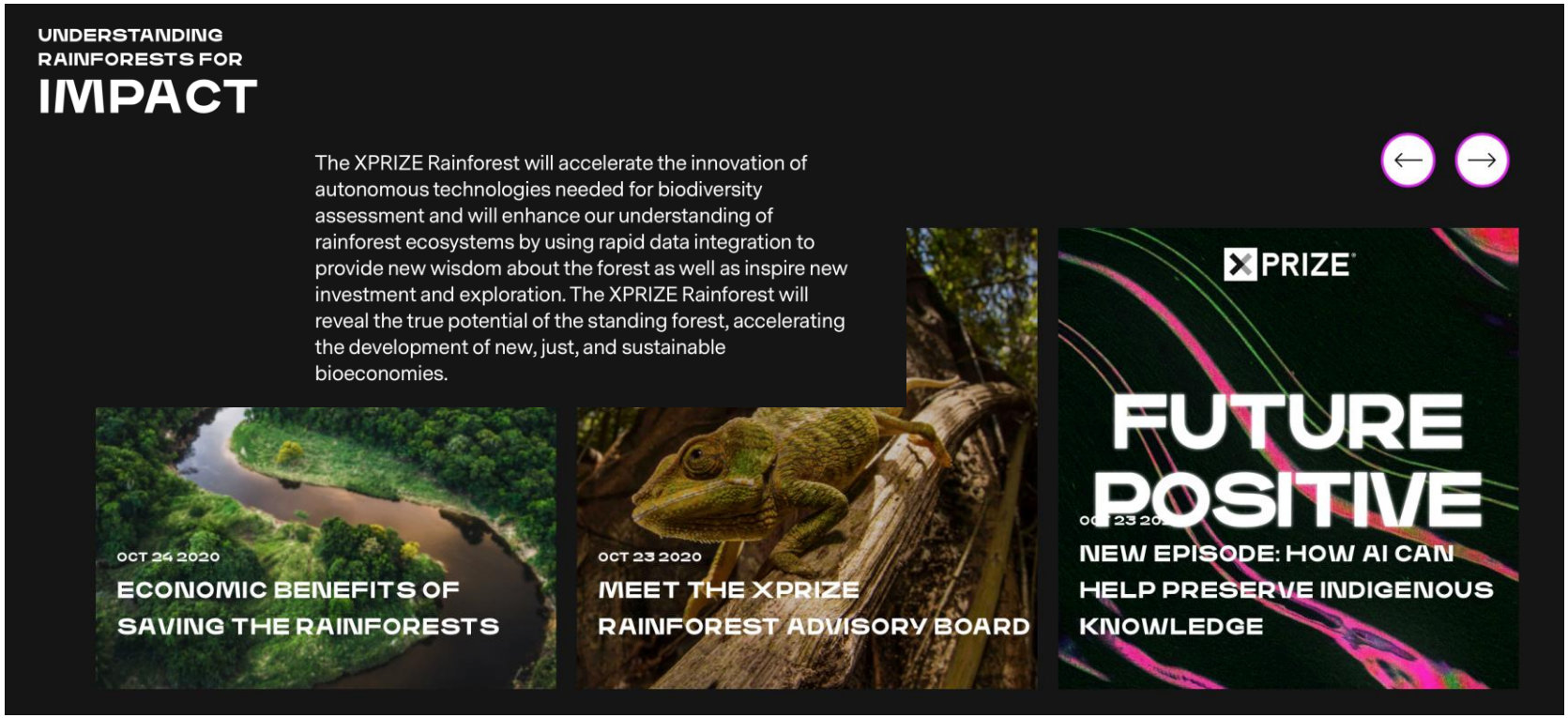






- Challenges: 1. New ways of working
    - High quality digitization of all material
        - A global digitized collection
        - Who will pay for it?
    - Sequencing of new material on arrival
    - First pass molecular/image-based/AI based classification
    - Efficient but accessible storage of physical material
        - What should be collected?
            - What do we mean by “reference” material?
        - How will collected material be used?

- Challenges: 2. Multi-scale data integration
  - Increasing large-scale, real-time analysis of the environment (remote sensing, metagenomics, etc.)
  - How can we ground these data back into reference material?



- Challenges: 3. Equitable benefit sharing
  - Increasing disagreement around implementation of the Convention on Biological Diversity, especially over the sharing of sequence data
  - Failure to reach agreement will likely lead to decreased sharing of materials and information, and decreased benefits for humanity in richer and poorer countries alike
  - Issues:
    - “Information wants to be free” versus property rights
    - The most biodiverse countries are generally poorer, and will incur costs if they they are to preserve biodiversity
    - “Benefit sharing” – but benefits can be long term and not easily monetizable

Our mission:

To understand and protect plants and fungi for the well-being of people and the future of all life on Earth

Five Scientific Priorities

- **PRIORITY 1:** Ecosystem Stewardship
- **PRIORITY 2:** Trait Diversity and Function
- **PRIORITY 3:** Digital Revolution
- **PRIORITY 4:** Accelerated Taxonomy
- **PRIORITY 5:** Enhanced Partnerships

Acknowlegements

PAFTOL *Bill Baker*, *Felix Forest*, *Ilia Leitch*

Darwin Tree of Life *Ester Gaya*, *Ilia Leitch*, Mark Blaxter Peter Hollingsworth. Michelle Hart Alex Twyford

Digitisation *Alan Paton*, *John Adcock*, *Sarah Phillips*, *Eduardo Toledo*, *Rob Turner*

Calleva Foundation

Department for Environment Food & Rural Affairs