Indexing for Life

Deliverable 2.6. Workshop 2 Policy Report: i4Life Bar-coding Workshop report

Work package 2

Editors: Vincent Robert, Alastair Culham and Magda Sitko

Contributors: Gianluigi Cardinali, Guy Cochrane, Alastair Culham, Mehrdad Hajibabaei, Pete Hollingsworth, Jean-Yves Rasplus, Sujeevan Ratnasingham, Stephane Riviere, Vincent Robert, David Schindel, Magda Sitko, Vincent Smith, Pelin Yilmaz

31 October 2012

Capacities Programme of Framework 7: EC e-Infrastructure Programme – Virtual Research Communities - INFRA-2010-2

Grant Agreement No:	261555
Project Co-ordinator:	Dr Alastair Culham
Project Homepage:	http://www.i4Life.eu
Duration of Project:	36 months
Start Date:	November 2010
End Date:	November 2013

i4Life Barcoding workshop report

European Bioinformatics Institute, Welcome Trust Sanger, UK. 24th - 25th September 2012.

The i4Life Barcoding workshop took place at the European Bioinformatics Institute, Hinxton, UK, on 24th afternoon and 25th morning September 2012.

List of participants:

Dr. Gianluigi Cardinali, gianlu@unipg.it
University of Perugia, Perugia, Italy

Dr. Guy Cochrane, cochrane@ebi.ac.uk
EBI-EMBL, Hinxton, United Kingdom

Dr. Alastair Culham, a.culham@reading.ac.uk
Reading University Herbarium, Reading, United Kingdom

Dr. Mehrdad Hajibabaei, mhajibab@uoguelph.ca
University of Guelph, Guelph, Ontario, Canada

Dr. Pete Hollingsworth, p.hollingsworth@rbge.ac.uk
Royal Botanic Garden Edinburgh, Edinburgh, United Kingdom

Dr. Jean-Yves Rasplus, rasplus@supagro.inra.fr
INRA, Montpellier, France

Dr. Sujeevan Ratnasingham, sratnasi@uoguelph.ca
University of Guelph, Guelph, Ontario, Canada

Dr. Stephane Riviere, sriviere@ebi.ac.uk
EBI-EMBL, Hinxton, United Kingdom

Dr. Vincent Robert, v.robert@cbs.knaw.nl
CBS-KNAW Fungal Biodiversity Center, Utrecht, The Netherlands

Dr. David Schindel, schindeld@si.edu
Consortium for the Barcode of Life, Washington DC, United States of America

Dr. Magda Sitko, m.h.sitko@reading.ac.uk
Reading University Herbarium, Reading, United Kingdom

Dr. Vincent Smith, vince@vsmith.info
Natural History Museum, London, United Kingdom

Dr. Pelin Yilmaz, pyilmaz.mgx@gmail.com
Max Planck Institute for Marine Microbiology Research, Bremen, Germany

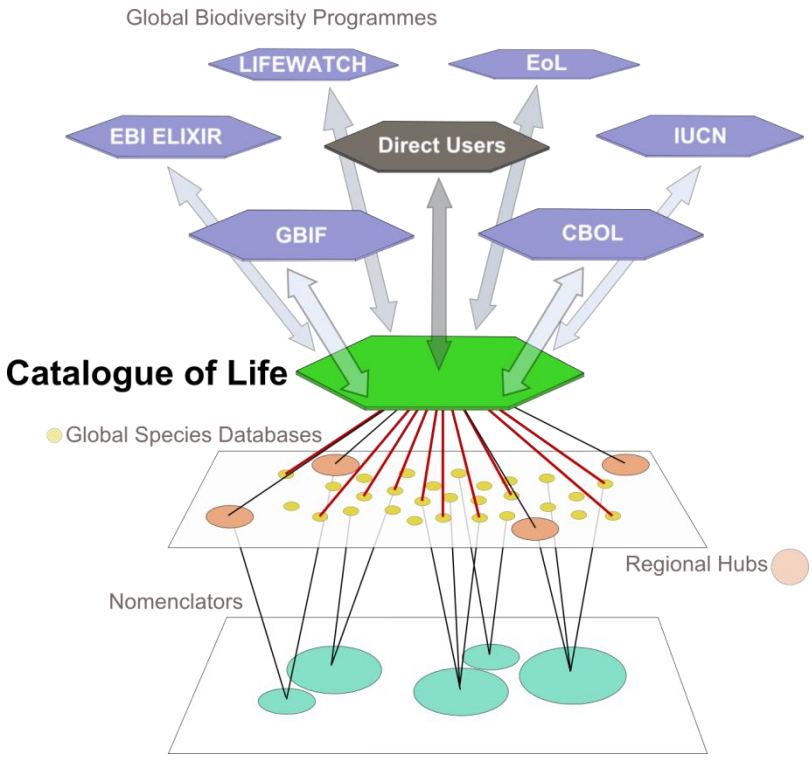
Schedule and summary of talks:

Monday 24th: Specimens/strains versus species

Opening and objectives of the workshop on barcoding by Dr. Alastair Culham

Dr. Alastair Culham (AC), the i4Life project coordinator gave a brief overview of the objectives of the i4Life project which is to establish a Virtual Research Community to interlink and harmonise global taxonomic catalogues. The existing Catalogue of Life (CoL) is used as a backbone. This builds on the work of the 4D4Life Project. AC gave a brief historic description of CoL and highlighted the tools that are available within CoL. CoL provides both dynamic and annual checklists, multilingual interface, scientific and common names, synonymy, distribution, references and scrutiny. i4Life project partners are some of the major global programmes (GBIF, EMBL-EBI, Barcode of Life, IUCN Red List, LifeWatch, Encyclopedia of Life, Sp2000, IT IS, University of Reading, ETI Bioinformatics, Cardiff University, MNHN Paris) exploring the full extent of life on Earth. This project will provide a summary of all species known across these programmes and create a global standard for taxonomic data integration in electronic infrastructures world-wide. AC described the WP of the i4Life project in details. Finally, AC gave the objectives of the current workshop:

1. To what extent are groups of sequences recognisable as taxa?
2. Can they have a stable name?
3. Should they be listed in Catalogue of Life?
4. Can i4Life begin a process of inclusion and should this be recommended?



Should we be working with specimens, strains, genomes, single sequences or continue with species? by Dr. Vincent Robert

Dr. Vincent Robert (VR) used to be the curator of the yeast collection of the CBS-KNAW, the largest culture collection in the world. He knows the fungal problematic very well and he is the head of a bioinformatics group of 13 researchers and software developers. His talk highlighted the identification problematic. He showed the history of methods and techniques used to identify fungi until now and demonstrated that morphology, physiology or other methods not based on molecular methods were usually not working on microbes. He explained why identification at species level remains important to determine the potential of unknown organisms. He also demonstrated the strong limitations of the traditional species centric approaches and indicated a number of examples where identifications might better be done at strains, specimens, genomes or sequences levels. VR also showed a number of yeast and macroscopic fungi examples where barcoding data can be used to reclassify wrongly named species. The Candida genus example was quite striking since one third of the Ascomycetous yeasts species are classified in the genus Candida while most of them belong to other existing or new genera. This highlighted the poor reliability and circumscription of microbial taxa. VR suggested that Catalogue of Life should not only include formally described families, genera or species but also references to non-described clusters of individual organisms. BINs (see below for more information) created on the basis of a given algorithm could represent such non-described groups. The fitting between existing and formally described taxa and BINs should also be reported. VR also suggested creating a platform allowing next generation sequencing data to be analyzed in a fast and accurate way. Current systems have serious scaling problems and new algorithms need to be developed to allow massive routine identifications and classifications using NGS.

Species delimitations and meaning for the identification of specimens or strains by Dr. Gianluigi Cardinali

Dr. Gianluigi Cardinali (GC) started to discuss on the existence of species and described several approaches for the circumscription and the dynamic of species: realistic, nominalistic, static or evolutionary. GC highlighted the various definitions which have been given of the term species and stated that no one definition has as yet satisfied all naturalists. GC cited Charles Darwin: "… yet every naturalist knows vaguely what he means when he speaks of a species. Generally the term includes the unknown element of a distinct act of creation". GC then pointed out that higher organisms species concepts are not necessarily adapted to the microbial world and he questioned whether the species level is the right unit for the study of microbial diversity. He also discussed the problem of finding discontinuities between groups of organisms. GC demonstrated that biological species concepts cannot be proposed as a mean to find discontinuities in asexual, parthenogenic or selfing forms organisms. GC showed the difficulties associated with quest for discontinuities in the character distribution. Differences within the species should be smaller than differences between species; differences between extreme individuals of two close species should be larger than differences between these and other individuals of the same

species. Well defined species should show a distance "jump" close to the species boundaries. A number of microbial examples were provided. For GC, microbial species are groups of strains artificially separated from the continuum of microorganisms due to the lack of a reproductive barrier system and species are different in size. GC concluded by stating that:

1. Microbes do not show significant discontinuities at the expected species level. It is likely that they are a sort of continuum, especially considering that no more than 5% of the total biodiversity has been isolated and described
2. This lack of discontinuities can be likely due to their non-obligatory sexual way of reproduction, which in turn is allowed by their unicellular nature. More complex forms of life have a different reproductive system, present reproductive barriers and are apparently a sort of discontinuum.
3. The nature of the species cannot be studied anymore without the contribution of the microbiological point of view.
4. Essential standardized principles should be established to produce algorithms for the classification and identification of microbes in a stable way.

The arthropods point of view by Dr. Jean-Yves Rasplus

Dr. Jean-Yves Rasplus (JYR) explained original goals of DNA barcoding that are to create an exhaustive catalogue of DNA sequences with two main aims:

1. Improve taxonomic knowledge of biodiversity with fast discovery of new species and creation biodiversity inventories
2. Species assignation of sampled individuals (whatever the sex and the development stages)

JYR gave some statistics on the current state of insect's barcoding. 170000 DNA barcodes of insect of 15000 species are currently available. It is estimated that 98% of insect species have no barcodes leading to frequent Type II errors (misidentification of queries without conspecifics in the database). Identification using COI (locus used for the barcoding of animals mainly) is globally acceptable but species delimitation needs more attention. JYR also showed a few examples of mitochondrial genes fragments incorporated to the nuclear genome that can be considered as pseudo-genes. The latter are co-amplification and can lead to over-estimation of species diversity. JYR also gave a few examples of Type I errors (misidentification of strains for which species are represented in reference databases) and discussed its possible sources. For JYR, the most obvious ones are:

1. COI sequence does not allow for the discrimination of some closely related species
2. Misidentifications
3. Contaminations
4. Incomplete lineage sorting
5. mtDNA introgressions

The danger of basing species concepts on one locus is clear and there could be advantages in using one or more nuclear genes to complement COI. A critical point is that sequences should be associated with vouchered specimens to help in the

reassessment of species concepts in case where used loci are confusing or misleading. JYR also discussed the limitations and advantages of current analytical methods such as sequence similarity, statistical classification, phylogenetic methods, population Models. JYR ended by suggesting to use sub-specific levels when needed, to use additional nuclear markers when necessary, not to change taxonomy unless using an integrative approach combining nuclear markers, morphological, biological and ecological data.

The plants point of view by Dr. Pete Hollingsworth

Dr. Pete Hollingsworth (PH) is a specialist of plant taxonomy and was the first author of the paper indicating which loci should be used for plant barcoding. Unlike in animals where the COI locus, the two selected markers for plants Rbcl and matK are not a providing sufficient resolution to distinguish many closely related species. Rbcl and matK have much lower discriminatory power than animal CO1. For PH, sequence-based clusters give most powerful insight into the distribution of plant diversity and provide the baseline framework for subsequent annotation of names and, in plants, 'Terminal' discontinuities are 'the norm' (and not sample density artifacts). Barcodes can also outperform binomials when characters are poor or in poorly studied groups where dispersal is good and hybridization levels are low. On the other end, binomials are sometimes more powerful than barcodes. PH indicated that sequence based clusters under-estimate plant diversity and have a lower information content in most circumstances. There are relatively few (genuine) discontinuities in the data, particularly at terminal nodes and barcodes are frequently 'shared' between species. There are fewer clusters than names and there may be operational difficulties in establishing a robust 'BIN' system for plants. Sequence based 'cryptic species' discovery will make a modest contribution to plant diversity. While the use of ITS locus alone is not possible, ITS and plastid can help resolving some clades but with unlinked multiple loci, incongruent phylogenies will be the norm. PH believes there is a need for mechanisms for detecting discontinuities in the data as a work-bench tool but clusters produced by such tools are likely to be unstable to additional sampling. He also suggests the need for the development of tools to identify character based differences separating closely related species. As with other groups, it would be useful to have a measure of identification certainty and annotation mechanisms to flag identification errors in public databases.

Much lower discriminatory power than animal CO1

– Floristic sampling:		
South Africa KNP	90%	Lahaye et al. (2008) PNAS **105,** 2923-2928
BCI Plot Panama	98%	Kress et al. (2009) PNAS **106,** 18621-18626
Amazonian trees	70%	Gonzalez et al. (2009) PLOSOne **4,** e7483
– Taxon based sampling:		
Inga	69%	Hollingsworth et al. (2009) MER **9,** 439–457
Araucaria	32%	Hollingsworth et al. (2009) MER **9,** 439–457
Asterella	90%	Hollingsworth et al. (2009) MER **9,** 439–457
Crocus	73%	Seberg et al. (2009) PLoS ONE **4,** e4598
Hordeum	50%	Seberg et al. (2009) PLoS ONE **4,** e4598
Carex (Canadian spp.)	95%	Le Clerc Blain (2009) MER **10,** 69 - 91
Acacia (6/6 spp)	100%	Newmaster et al. (2009) MER **9,** (Suppl. 1), 172–180
Berberis	23%	Roy et al. (2010) PLoS ONE **5,** e13674.
Palms	81%	Jeanson et al. (2011) Annals of Botany
Picea	25%	Ran et al. (2010) J. Int. Plant Biology **52**, 1109-1126
Taxus	100%	Liu et al. (2010) MER **11,** 89-100
Fabaceae	80%	Gao et al. (2011) Planta Medica **77,** 92-94
Solanum	12%	Spooner (2009) A JB **96,** 1177-1189

Tuesday 25th : New tools and technologies

BIN system for the automated clustering/Identification of OTUs by Dr. Sujeevan Ratnasingham

Dr. Sujeevan Ratnasingham (SR) gave a quite complet overview of the Barcoding of Life Database (BOLD) and described the existing and newly developed tools for the management and the analysis of barcoding data. He also showed the large acceptance and interest in barcoding as well as the huge increase in data production since the start of the project. He pointed out the difficulties encountered with data release policies and issues related to data quality and annotations. SR and his team created a new method for the automated discovery and creation of clusters called Barcoding Index Numbers (BINs). For animal barcodes (COI), it seems that there is concordance with existing species circumscriptions, that the BINs system is stable and persistent. He stated that the system can handle singletons and supports third party annotations. SR showed some clustering results obtained using several methods/algorithms like jMOTU, ABGD, CROP, GMYC and BIN. BIN performed usually better except for ABGD which had comparable results. Most clusters or BINs produced are singletons and variation does not go up with sampling.

NGS and its impact on Barcoding and species descriptions by Dr. Mehrdad Hajibabaei

Dr. Mehrdad Hajibabaei (MH) presented the state of the art of the Next Generation Sequencing (NGS) methods. MH showed that while the known biodiversity is around 1.9 million species of plants and animals while the estimated biodiversity is in the range between 10 million to 100 million species. Unfortunately, limited sampling of the world's biodiversity to date has prevented a direct quantification of the number of species on Earth, while indirect estimates remain uncertain due to the use of controversial approaches. MH gave a number of examples where NGS can be applied such as biodiversity studies, human microbiome, macrobiotic studies, biomonitoring, ecological studies, etc. While DNA sequence information have extensively been used for biosystematics including Tree of Life and Barcode of Life. MH believes that different genes are required to provide resolution at different taxonomic levels and domains of life and that one gene does not fit all. The key for monitoring applications versus one-off studies is how to access DNA information. Current approaches based on single specimens are not scalable. The major

challenges of the future are the bioinformatics aspects: storage, management and analysis of data.

CBOL and Barcoding general perspectives by Dr. David Schindel

Dr. David Schindel (DS) is the leader of the Consortium for the Barcode of Life (CBOL) that is responsible for the establishment of standards associated with barcoding. DS described the rationale behind all the information required when depositing barcoding sequences in one of the BOL databases (taxonomic identification to species, voucher specimen ID in standard format, name of barcode region, country/ocean/sea of origin, latitude/longitude, name of collector, collection date, name of identifier, length, quality, two trace files and forward/reverse primer sequences, names). New required data will soon be required such as: taxonomic reliability, type status of voucher specimen, basis for identification, type comparison, subjective confidence level for identification.

DS distinguishes three types of barcoding initiatives:

1. project based, which, for DS are the most interesting scientifically
2. collection based, most cost-effective and having support of institutions
3. application based, more sustained funding

He also spoke about challenges such as data release problems (only a limited portion of barcode data are publicly available), compliance with data standards, species delimitations, long-term management and curation of barcode databases, etc. CBOL is trying to promote DNA barcoding as a global standard for species identification and as such would be happy to see the outcome of barcoding more vused and visible in systems such as the CoL.

General discussions and conclusions

Two general discussion sessions have been organized at the end of each working day and the main conclusions are the following:

1. For microbes, traditional species concepts are usually not well adapted for use, and working at strain/specimen level would probably be more accurate. Use of dynamic clustering would also be a plus since it would create groups according to the interest of the users of the data rather than according to taxonomic placements only. Also, and unlike animals and plants, most of the diversity remains to be discovered. Therefore, NGS and environmental sampling are likely to provide huge amounts of new taxa that won't be vouchered and properly described using traditional classification methods. Automated clustering or BINning might be of great importance in such a case.
2. Situations are quite different depending on the organisms (microbes, animal, plants) and the barcode regions selected for the latter, but as a general rule, a multiple loci approach will be much better than single locus sequencing for both

the identification and classification steps. Also, for the automated clustering or BIN system, one single gene will not fit all problems and there could be serious resolution problems (at least in microbes and certainly in plants) if one uses only the current official barcoding regions as a basis.

3. If using the BIN system or similar implementations, there should a way to measure the fitting level between automated and manual classifications. Also the methods and data used for the automated BINing should be clearly stated on the websites displaying BIN numbers.
4. The automated BINing system should not replace classical taxonomic placements or classifications but should rather complement it, especially for unknown clades.
5. BIN numbers should be as stable as possible and traceable.
6. Barcoding standards are regarded as very useful and should be required as much as possible, there is a serious problem with environmental sequencing where, by definition, no specimen vouchers can be given. The participants recommend gathering as much data as possible and establishing standards for environmental samples as well.
7. While some people believe that manual annotation of data/records could be a nice solution to curate existing deposits, others seriously doubt that this would be an efficient and scalable system for the management and curation of massive databases that will inevitably be created as a result of environmental sequencing.
8. It is felt that large central databases might play a valuable coordinating role but that distributed, specialized databases with many more metadata related to the specimens, sequences, their ecology and biology will be essential to extract the full potential from data gathered. It seems that metadata might become at least as important as sequence data in the future.
9. Implemented systems need to be appropriate and useful to the end-users otherwise there is not much point to include BINs at this stage.

Short answers to the specific questions or goals of the barcoding workshop:

1. To what extent are groups of sequences recognisable as taxa?

As mentioned above, there is certainly not always a very good fit between actual taxa and sequences.

2. Can they have a stable name?

Like traditional classification taxonomic work, names might not be completely stable. As the number of samples/specimens/sequences is growing, it will inevitably create situations where the automated classifications or previous taxonomic denomination of sequences will have to be adapted or changed. For microbes and possible other groups, “Candidate BINs” should be created. So depending on the taxonomic groups, there might be different rules for the creation of BINs.

3. Should they be listed in Catalogue of Life?

We all feel that barcoding-associated taxonomic data should be incorporated into the CoL once sufficient data are there to allow some stability. This would help estimating the amount of diversity not captured with traditional taxonomy.

4. Can i4Life begin a process of inclusion and should this be recommended?

Yes, we recommend that such data should be included in CoL. If resources to do so are available within the timeframe of the i4Life project, this process could begin providing the Catalogue of Life Directors agree. However, this would require moving already committed resources from other activities and depends on completion of those activities ahead of schedule and below budget.