

Assembling the Tree of Life - Europe (ATOL)

Report from a Meeting held at
Royal Botanic Gardens, Kew on 14th and 15th October, 2002

Summary

Assembling the Tree of Life - Europe (ATOL) is an Integrated Project aiming for funding in EC FP6. ATOL will produce the ultimate taxonomic and phylogenetic resource for biological research and general use: a Tree of Life (ToL) database. New data will be generated on a massive, coordinated scale by linking European taxonomic expertise and Biodiversity collections to molecular lab and bioinformatics expertise. A backbone for the global ToL will be established in collaboration with global initiatives, together with detailed studies on target taxa with high impact on society and the environment. A task force will target taxa for which there is taxonomic expertise and need but no current phylogenetic research. A network of local lab centres will be set up to coordinate lab methodology and training within the project. A European Phyloinformatics Centre will be created as a high-prestige, visitor-driven research centre to curate the ToL database and develop analytical tools for reconstructing the ToL. Key actions needed for the proposal are: to work out details of teams and topics/taxa to be included, to work out business opportunities in greater detail, to set up a management structure for future stages that combines managerial experience with 'hands-on' enthusiasm.

Introduction

1. On 10th May 2002 Diana Lipscomb Program Director of the US NSF Tree of Life (ToL) Initiative sent an e-mail to a number of systematists in Europe encouraging them to submit a ToL Expression of Interest (EoI) to the European Framework 6 (FP6) funding round. As a result Vincent Savolainen and Tim Barraclough sent a circular e-mail proposing to submit an EoI for an Integrated Project (IP). By 6th June 2002 membership had snowballed and we submitted an EoI to the Thematic Priority 1.1.6.3. 'Global Change and Ecosystems'. The consortium included 45 institutions from 14 countries and has since grown. Background information on FP6 is available from Cordis (www.cordis.lu) and, among others, the UK Research office (www.ukro.bbsrc.ac.uk).
2. A meeting was held at The Royal Botanic Gardens, Kew on the 14th and 15th of October to plan strategy and subsequent stages of the proposal. 55 people attended from 15 countries, plus Diana Lipscomb from the US NSF and Annette Hayden and Damian Walsh, EC advisors from Imperial College Research Contracts office. Day 1 comprised presentations from a diverse collection of participants, Day 2 comprised group discussions on components of the project. The meeting program and list of participants is in the appendix.
3. We thank Mike Bennett (Royal Botanic Gardens, Kew) and Martin Buck (Dept. Biological Sciences, Imperial College) for financial support, and all the participants for funding their own travel and accommodation.
4. This report summarises the main points discussed in the meeting and highlights action points needed for the next stages of proposal and planning. **It should be read in conjunction with the EoI and meeting program.** Please send feedback on any issue to Vincent Savolainen (v.savolainen@rbgkew.org.uk) or Tim Barraclough (t.barracough@ic.ac.uk).

General

5. We have had excellent feedback on the ATOL EoI. The scale of ambition and breadth of membership has been judged to be excellent by several EC funding advisors. The scientific plan and broad strategy outlined in the EoI was largely upheld by participants at the meeting, with some exceptions outlined below.
6. **We received feedback at the meeting and elsewhere that ATOL would feature in one of the calls for full proposals**, despite no explicit mention of ToL topics in the call for EoIs or some subsequent on-line material. **This can still change before the full call.** We will know for certain by Nov/Dec, but for now assume that ATOL will submit a full proposal.
7. **We recommend that the ATOL EoI and this report be used as a starting point for any full proposal.**
8. ATOL membership has mostly good taxonomic and regional coverage but there are relatively few women.

Phyloinformatics

European Centre for Phyloinformatics

9. There is general consensus that setting up a Phyloinformatics Centre should be a key part of the ATOL project. The Centre would: curate the data arising from the project; develop, maintain and implement new methods, software and hardware needed for data analysis; develop database and web systems for presenting the data for research and general end-users, coordinating with any future initiatives elsewhere in the world; act as a strong centre for phyloinformatic research in Europe and beyond.
10. The proposed model is a centre headed by a director, with support staff and resident researchers, but populated largely by visitors comprising academics, post-docs and graduate students from around the ATOL network and beyond. The Isaac Newton Institute in Cambridge was raised as an example. This model would provide a strong, high prestige focus for research across Europe and very strong training. **Numbers of resident researchers, support staff and visitors need to be decided.**
11. Communication between the Centre and the organismal teams should be a priority. Data acquisition strategy should have input from bioinformaticians, and computer strategies should have input from organismal teams. Visitors to the Centre would include researchers from the taxon teams as well as bioinformaticians.
12. ATOL needs to guarantee the lifespan of the Centre and databases beyond the project funding. There was general consensus for linking the Centre to existing major databases within Europe, such as the EBI. Thus, Cambridge, Heidelberg and Lyon were suggested as possible locations. **We need to decide a preferred location, contact the relevant institution, and work out a financial plan to ensure longevity of the Centre.**
13. Phyloinformatic projects were not successful in the first round of USA ToL. We have a unique opportunity to play a lead role in global initiatives. **ATOL should contact leaders of the unsuccessful bid to the NSF ToL to coordinate our efforts and ideas.**
14. Computer science expertise within ATOL is biased towards methods and raw software production. Although the computer science needed to develop a polished end resource is not sophisticated, **greater named expertise in software engineering, database and web design is needed for the proposal stage.**
15. Many details need to be worked out by the likely call date. **A team should be set up immediately to work on Phyloinformatics details of the proposal.** We propose the

following: Rod Page (chair), Olivier Gascuel, Michel Milinkovitch, Andrew Rambaut, James McInerney, Fredrik Ronquist, Mark Wilkinson, Paul Agapow, Mark Pagel, David Liberles and Olaf Bininda-Emonds. We stress that this is not the final bioinformatics team, just a group to work out details for the proposal. Contact us to suggest extra members.

Resources and analysis.

16. The general consensus is for ToL data to be made available as an on-line resource, with the goal of publishing a 'working draft' of the Tree of Life by 2008/2010. This is a good deliverable but we need to think hard about what exactly we can deliver by what time.
17. **We need to outline what data will be provided by the final resource.** The minimum is trees and the underlying data they are based on. Timing of events on the tree in millions of years would also be highly desirable for many likely end-users.
18. Based on the EoI and participant views, ATOL will focus on DNA sequence data. **The use of morphology and other data needs to be clarified: should they be included where possible or only where sequence data are lacking/deficient?**
19. As well as being a repository of phylogenetic data, the ToL database would provide a 'best estimate' of the ToL that could be queried by end-users wanting phylogenetic information for a group, parts of the tree lacking current data etc. To be a research tool, the underlying data must be downloadable in useful form e.g. Nexus files. One possibility is for different levels of entry for different end-users: a user-friendly pictorial system for non-research users, but more detailed front-end for researchers. **Query and web design needs to be worked out in detail by the next proposal stage.**
20. Submission of data to the ToL database would be open to the whole scientific community, not just members of ATOL. The database would be free and open. The model best suited to this goal is on-line submission of data, similar to that used by Genbank. **A plan for curation of the data is needed, to maintain quality of input and therefore output.**
21. **Algorithms for integrating online submission with data analysis to produce the best estimate ToL need to be outlined in the proposal.** Bayesian methods were presented as a way to feed in new data without needing to restart computer searches. Supertree methods are critical for producing a final best estimate of the ToL. Should we provide tools for users to conduct their own analyses, including hardware resources for complex searches, or simply provide results from one or a few methods of analysis? **The steps from data to ToL need to be finalised, driven by best practise and user need.**
22. **We need to outline how we will deal with and represent uncertainty in the final product.** Uncertainty should be apparent at all levels of entry e.g. different colours to branches with low support or based on different quality or quantity of data. We need to decide approaches for groups like bacteria that might lack a true single tree. Very basal branches in the tree of life might be affected by similar problems.
23. No discussion emerged of possible links to Small-Medium Enterprises. The data should be free and open, but analytical and/or presentation services could be provided to industry. Existing SMEs could be used to provide software engineering, web design etc.
24. **Topics needing research as part of ATOL must be outlined and teams assembled.**

Phyloinformatics database

25. **We stress that storage of the data in a flexible format is really the key issue.** 'Best practise' of analysis might change in future, but if the data is stored in the right format, analysis functions can be easily changed.
26. ATOL will initiate major data collection, but automated data mining of existing sources of phylogenetic data is also needed: we do not want to start from scratch. **The proposal should outline methods and legalities of trawling Genbank, online journal archives etc. to transfer data into the ATOL database.**
27. Different models of data storage were presented. The model that fits best our goals and IP funding expectations is a database kept on disk space at the Phyloinformatics Centre and a GRID facility for conducting analyses. The database would communicate with other global centres as and when they arise. We should plan capacity for a 10-100X increase in the rate of data collection.
28. Compiling multiple datasets and presenting a ToL requires consistent use of species and higher-taxon names. At present no single resource exists to provide a globally accepted set of names for a taxon. **We need to outline our solution to the naming problem,** for example feeding our names into existing species name databases such as GBIF **coordinating with other submissions to FP6 on this topic.**
29. A directory of researchers listed by taxon and genes/characters would be useful.
30. Suggestions were made to incorporate additional data into the database, e.g. morphology and genomics, to facilitate comparative biological research such as EvoDevo. This raises problems of where to stop (ecological data etc.) and also of data quality: such data are likely to be very incomplete and of different kind among different taxa. **We recommend that these kinds of data are linked to ATOL as a parallel database rather than integrally, developed initially by target taxon teams or for a few case study taxa.**

Phylogenetics and data generation

Strategy

31. ATOL should adopt a two-pronged approach to generate new phylogenetic data: i) establish a ToL framework or backbone, ii) focus on target groups with justifiable impact on society or the environment. **The sample should be global rather than European,** but driven by the existence of European expertise and access to samples.
32. The main aim and strength of ATOL is to go beyond existing research to a new level of coordination. At all levels, ATOL must be greater than the sum of its parts: it is not simply a collection of smaller taxonomic projects. Emphasis is on coordinating efforts to achieve ambitious goals of completeness for the ToL backbone and target taxa and to empower new research to fill in gaps in current research effort.
33. **We need to state carefully what our deliverables will be at the end of the funding.** It was pointed out that a 100% complete ToL backbone even to phylum is probably not deliverable due to the extreme rarity of some taxa.
34. For the target groups, the level of study will be group dependent. Mostly we should stress completeness (+/- rare taxa) at the level of family or genera for a higher group, but with some case studies at species-level looking towards the future of ToL initiatives.
35. There seemed to be a majority view that DNA sequence data should be the main focus for tree reconstruction in ATOL, but **with morphological and non-sequence molecular data playing a role that still needs to be clarified.** We propose that teams working on target groups should aim principally for a DNA sequence tree but with members working on other data sources where necessary/interested.

36. Linking taxonomic expertise with molecular and phylogenetics expertise is the key to the success of ATOL. The lasting impact of ATOL will be to gather a collection of well-identified specimens annotated with information on DNA and other attributes. **We need a very clear strategy for how to engage taxonomists' involvement, how to fund or pay for their efforts, how to guarantee supply of samples within legalities.**
37. The third aim outlined in the EoI, gap-filling, is not scientifically distinct but could be retained as a separate 'work package' for strategic and logistical reasons. **ATOL should establish a task force to set up phylogenetic work on taxa for which there is taxonomic expertise but at present no phylogenetic studies.** This would be strong component under expectations for an IP.
38. Maintaining a tissue collection for all material processed by ATOL is essential. This would be an unparalleled resource for future studies on DNA, RNA etc. Tissue collections are being set up in several European biodiversity centres: ATOL should reach agreements with these centres, and drive for unified procedures, computer databases etc.
39. **We need to establish teams and team leaders as soon as possible.** This process should be open to researchers outside the present ATOL membership. A Project Management Committee (PMC) will be set up to manage this process. **We call for 1-2 page outlines of proposed teams and target taxa, preferably from groups of people but also from individuals who might fit into final teams. The philosophy of ATOL is to coordinate research across Europe rather than run a set of parallel but uncoordinated projects. Hence the PMC will choose teams and topics based on the big picture, and actively canvas topics and researchers where necessary.** ATOL will aim to choose target taxa different from those funded in the first round of the USA NSF ToL initiative. Submissions will be reviewed by the PMC. An open call will be made once we know if ATOL is included in the call for funding.

DNA sequence data

40. Teams working on the ToL backbone and on target taxa should aim to generate DNA sequence data for at least two, ideally more, DNA regions spanning multiple linkage groups. Markers need to be consistently amplified across the breadth of the taxon. For target taxa, choice of marker should be left to the taxon team: supertree methods will be used to piece together the wider tree. For the ToL backbone, a team should be assembled to coordinate efforts and to develop appropriate markers. The Phyloinformatics Centre and team should be involved in planning stages.
41. Expressed Sequence Tag (EST) approaches were presented as tools to develop new markers appropriate to particular levels of study. Production of cDNA libraries and microarrays for representatives of most major taxa was highlighted as a side product with huge business and research potential: a phylogenomic resource. Central facilities should be set up to train people in these techniques and to optimise methodology across teams.
42. Whole mitochondrial DNA sequences offer an alternative source of data. Any efforts in ATOL should collaborate with established programs in the USA and Australia. ATOL should investigate the use of molecular synapomorphies, particularly for deep branches.
43. DNA sequencing should be kept 'in house' to retain control over pricing etc. However, one possibility would be to set up an SME to provide sequencing for ATOL but after funding stops to offer a specialist sequencing service for taxonomic/biodiversity end-users.
44. **The PMC needs to work out a detailed, coherent description of sampling and lab strategy for the proposal as soon as possible.**

Central facilities

45. **We need to decide immediately what central lab facilities should be set up.** One suggestion was to have a Phylogenomics Lab Centre run parallel or integrated with the Phyloinformatics Centre: no details were discussed. An alternative was to have 3-5 regional existing labs designated as specialist centres for tasks common to the entire work program. E.g., there could be a centre for cDNA library construction. At least one centre should specialise on providing training and facilities to under resourced parts of Europe. Centres could provide sequencing facilities for work by ATOL teams. The centres would act to integrate lab work under ATOL, minimising redundancy in development and optimising techniques across the network. **Details including names and locations of labs need to be worked out.**
46. Priority would be to ATOL members, but centres might also provide business services to industry users, see below.

ATOL providing a research and business resource

47. ATOL should go beyond simply establishing and presenting a ToL, it should develop research and business uses as an integral part of the project. **Much greater thought is needed on business and SME opportunities and to detail names and numbers. Emphasis within ATOL should be placed on a few strong and very direct practical uses.**
48. DNA taxonomy and identification is an area with obvious business potential: a database with DNA sequences obtained from well-identified specimens can be used to offer identification services to agricultural, environmental and forensic end-users. **To pursue this we need to work out the taxonomic level of identification offered and what taxa would be targeted for this aim, e.g. soil organisms, microbes, pests, marine organisms etc.**
49. Phylogenomics is another area where ATOL should develop research and business potential. Possibilities include setting up an SME to provide cDNA/BAC libraries or microarrays for non-model organisms (a phylogenome service), providing analytical services for industry uses of ToL information, e.g. molecular adaptation, and developing databases to annotate the ToL with more detailed genomic data.
50. Evolution of development (EvoDevo) was another area raised by participants. It was suggested that a parallel morphological database might be established with a principle aim of facilitating EvoDevo research. Strategy for these areas needs to be worked out, but **we believe ATOL should facilitate rather than fund this area because of its weaker link to our strong central goal.** See point 30.
51. In the EoI we referred to 'function and evolution of diversity in ecosystems' as an integral research interest but this has not yet been discussed in detail. Given the Priority Area for our submission, **we need to develop a concrete plan on this topic.** Again, we think facilitation rather than strong funding should be the emphasis, with ATOL involvement distributed among the target taxon teams and Phyloinformatics Centre.
52. **Detailed plans of SME/business aspects need to be worked out.** Expectation overall for IP projects is that 15% of funds go to SME. Our topic is expected by the EC fall below this average, but still there is an advantage for maximising our SME potential. Possibilities are that centres or teams within ATOL provide business services or that we set up or collaborate with 'external' businesses.

Management

53. **A strong management plan is vital for the success of ATOL.** The proposed budget will be around 30 million euros. The coordinator is expected to be a high-profile scientist, project manager and finance manager. Access to administration, experience of this scale of project and EU funding are all important. The coordinator is expected to spend 50% of their time on the project.
54. At the same time, **a large part of the success and enthusiasm of ATOL has come from the roots up.** It is vital to keep this momentum at the science level.
55. **We propose the following model:**
- I) Imperial College (IC), Kew Gardens and the Natural History Museum will act as a host 'consortium' with one of them chosen as host institution. A senior scientist from each institution will be on the management committee, one of them as PI and coordinator. Together the institutions provide a very strong portfolio with respect to the above requirements. The precise role of each institution cannot be finalised until the model contracts for FP6 are published. This plan has backing from Profs. Peter Crane (Director, Kew), Mike Hassell and Charles Godfray (Head of Faculty and Director of Centre for Population Biology, IC) and the Natural History Museum.
 - II) The host institutions have agreed to fund a dedicated Project Manager for 12 months if the ToL is indeed listed in the call for proposals. The manager would run administration of the proposal stage. If the project is funded there would be a management office with manager, finance officer and secretaries etc.
 - III) A Project Management Committee will be set up that includes senior members of the ATOL network with managerial and scientific experience (including the 3 senior scientists from the host institutions and a senior finance officer) plus a set of researchers providing 'hands-on' scientific impetus on the key areas of ATOL. It is suggested that a senior US scientist should also be on the committee, and representative 'end-users' and an Intellectual Property expert. The committee will be representative in terms of study taxa, geographic area, and expertise.
 - IV) The Committee will manage all aspects of the proposal. Its first tasks will be to decide on the detailed management structure for the project (if funded) and to decide on the detailed scientific program guided by the EoI and meeting report.
 - V) Overview of general structure of ATOL is shown in the appendix.
 - VI) Details of this model and people involved cannot be finalised until after the call for proposals in Nov/Dec. We have a list of names based on suggestions and input into previous stages, but please send further ideas to Vincent or Tim as soon as possible.

Glossary

ATOL - Assembling the Tree of Life - Europe

EoI - Expression of Interest

EST - Expressed Sequence Tags, method for sequencing expressed DNA via RNA extraction

EvoDevo - evolution of development

FP6 - Framework 6, imminent European funding round

GBIF - Global Biodiversity Information Facility

IP - Integrated Project, new instrument of funding for FP6

Phylogenomics - phylogenetic scale information that goes beyond one or a few gene sequences

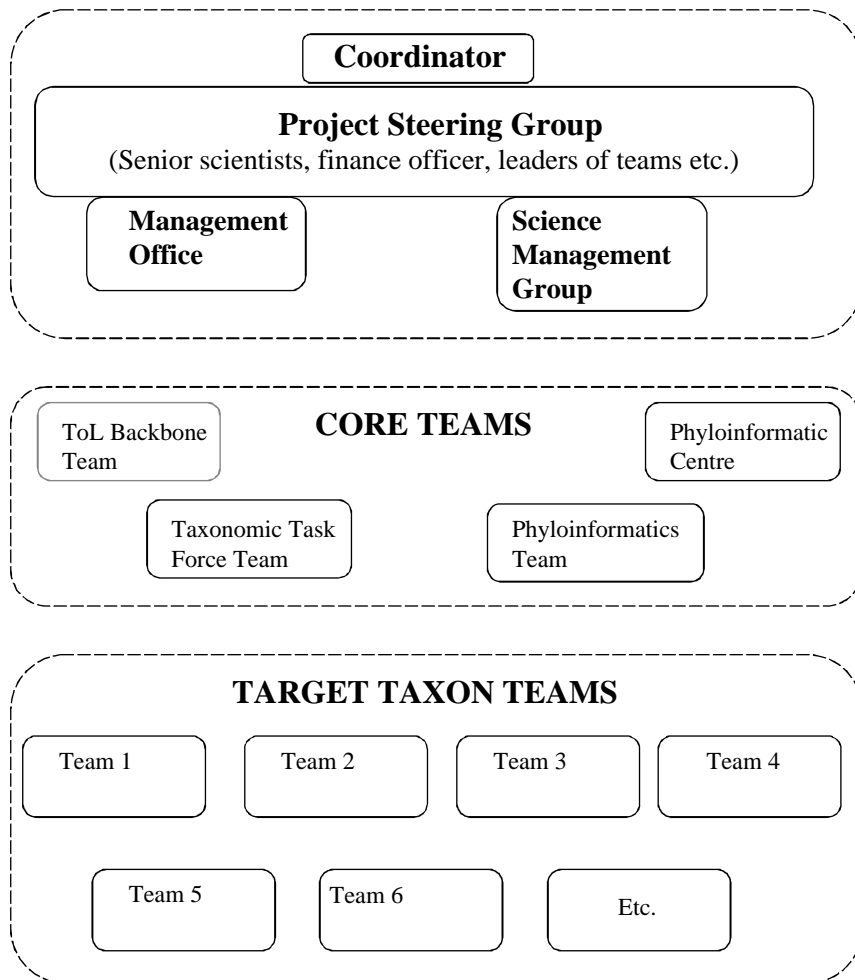
Phyloinformatics - Bioinformatics of phylogenetics, i.e. mathematical, computer aspects

SME - Small-Medium Enterprise

ToL - Tree of Life, i.e. evolutionary relationships among all organisms

APPENDIX

PROPOSED STRUCTURE OF ATOL



ATOL Workshop, 14-15 October 2002

Venue: Jodrell Laboratory, Royal Botanic Gardens Kew, UK

Organised: Vincent Savolainen (Jodrell Laboratory) and Tim Barraclough (Imperial College at Silwood Park).

Presentations

Angela Newton: bryophytes
Francesco Frati: arthropods
Rod Page: TreeBase
Vincent Berry: bioinformatics
Emmanuel Douzery: mammals
Michel Millinkovitch: large searches
James McInerney: methods
Mark Wilkinson: supertrees
Julie Hawkins: morphology/EvoDevo
Diana Lipscomb: ATOL USA
Donald Quicke: wasps
Vincent Savolainen: angiosperms
Alfried Vogler: DNA taxonomy
Sofija Pekic: agriculture/Belgrade
Piotr Bednarek: Eastern Europe
Lene Lange: Fungi/SMEs
Kosmas Theodorides: phylogenomics
David Liberles: phyloproteomics
Max Telford: molecular synapomorphies

Group discussions:

1. Bioinformatics (chair Rod Page)
2. Phylogenomics (chair Alfried Vogler)
3. DNA and beyond! (chair Julie Hawkins)
4. Management (chair Chris Humphries)

PARTICIPANTS

ALAN PATON	Royal Botanic Gardens, Kew, UK
ALFRIED VOGLER	The Natural History Museum, London, UK
ANDREW RAMBAUT	Dept of Zoology, Univ of Oxford, UK
ANETTE HAYDEN	Imperial College EU Advisor
ANGELA NEWTON	The Natural History Museum, London
ARNE NOLTE	Genetics Institute, Univ of Koeln, Germany
BERND SCHIERWATER	University of Hannover, Germany
BILL BAKER	Royal Botanic Gardens, Kew, UK
CEDRIC BERNEY	Dept of Zoology, Univ of Geneva, Suisse
CHRIS HUMPHRIES	The Natural History Museum, London, UK
CHRISTOPH OBERPRIELER	Botan Museum Berlin-Dahlem, Freie Univ Berlin, Germany
CHRISTOPHE THÉBAUD	University Paul Sabatier, Toulouse, France
DAVID LIBERLES	Stockholm University, Sweden
DAVID SPRINGATE	Royal Botanic Gardens, Kew, UK
DIANA LIPSCOMB	National Science Foundation, USA
DION DEVEY	Royal Botanic Gardens, Kew, UK
DONALD QUICKE	Imperial College, Ascot, UK
ELISABETH HERNIOU	Imperial College, Ascot, UK
EMMANUEL DOUZERY	Univ of Montpellier 2, Montpellier, France
FABIAN HAAS	Universität Ulm, Sektion Biosyst. Dokumentation, Germany
FRANCESCO FRATI	Dept of evol. biology, Univ of Siena, Italy
FRANÇOIS BALLOUX	University of Edinburgh, UK
FRED SCHRAM	Zool. Museum, Amsterdam, The Netherlands
FREEK BAKKER	Nationaal Herbarium Nederland
JAAKKO HYVÖNEN	University of Helsinki, Finland
JAMES MCINERNEY	Dept of Biology, National Univ of Ireland
JAN KRAIC	Div. Appl Genet. & Breed.Res. Inst. Plant Prod., Piešťany, Slovakia
JEAN MARIAUX	Nat. Hist. Museum of Geneva, Switzerland
JIM CLARKSON	Royal Botanic Gardens, Kew, UK
JOE BIELAWSKI	University College London, UK
JULIE HAWKINS	Univ of Reading, Reading, UK
KÅRE BREMER	Univ of Uppsala, Uppsala, Sweden
KOSMAS THEODORIDES	The Natural History Museum, London, UK
LAURE CIVEYREL	University Paul Sabatier, Toulouse, France
LENE LANGE	Novozymes A/S, Bagsvaerd, Denmark
LOUNES CHIKHI	University College London, UK
MARK WILKINSON	The Natural History Museum, London, UK
MARTYN POWELL	Royal Botanic Gardens, Kew, UK
MAX TELFORD	Univ. Museum of Zoology, Cambridge, UK
MICHEL MILLINKOVITCH	Inst. of Molec. Biol. and Medec., Free Univ of Brussels, Belgium
NEIL CAITHNESS	University of Reading, UK
OLE SEBERG	Dept of Evol. Botany, Univ of Copenhagen, Denmark
PAULA RUDALL	Royal Botanic Gardens, Kew, UK
PETER LINDER	Dept of Botany, Univ of Zurich, Switzerland
PETER OLSON	The Natural History Museum, London, UK
PIOTR BEDNAREK	Centre for Biol. Diversity & Conserv, Polish Acad of Sci, Warsaw
RAFAEL ZARDOYA	Natural History Museum, Madrid, Spain
RICHARD THOMAS	The Natural History Museum, London, UK
ROD PAGE	University of Glasgow, Glasgow, UK
SOFIJA PEKIC	University of Bgrade, Serbia
SOILI STENROOS	University of Turku, Finland
TIM BARRACLOUGH	Imperial College, Ascot, UK
TREVOR HODKINSON	Dept of Botany, Univ of Dublin, Ireland
ULRICH NUBEL	German Collection of Microorgs and Cell Cultures GmbH, Braunschweig, Germany
VINCENT BERRY	Univ of Montpellier 2, Montpellier, France
VINCENT SAVOLAINEN	Royal Botanic Gardens, Kew, UK