

SEQUENCING OF THE *EUCALYPTUS* GENOME: A PROPOSAL TO DOE-JGI



E. grandis tree. Photo: Z.Myburg

Prepared by

**Zander Myburg, Dario Grattapaglia, Brad Potts, Carlos Labate, Gerd Bossinger,
Margaret Byrne, Rene Vaillancourt, Ron Sederoff, Simon Southerton**

and other members of

**the *Eucalyptus* Genome Network
(EUCAGEN)**

18 June 2006

Summary

A major challenge for the achievement of a sustainable energy future is understanding the molecular basis of superior growth and adaptation in woody plants suitable for biomass production. For this reason, we propose the sequencing of the genome of a *Eucalyptus* tree. *Eucalyptus* species are among the fastest growing woody plants in the world (with mean annual increments up to 100 m³/ha). *Eucalyptus* is the most valuable and most widely planted genus of plantation forest trees on a world-wide basis (approx. 18 million ha) due to its wide adaptability, extremely fast growth rate, good form and excellent wood and fiber properties. Genome sequencing is essential to understand the basis of its superior properties and to extend these attributes to other species. Genomics will also allow us to adapt *Eucalyptus* trees for green energy production in regions (such as the Southeastern USA) where it cannot currently be grown. The unique evolutionary history, keystone ecological status and adaptation to marginal sites make *Eucalyptus* an excellent focus for expanding our knowledge of the evolution and adaptive biology of perennial plants.

1. Introduction

Eucalyptus species, commonly referred to as eucalypts, are native to Australia and the islands to its north. They are generally long-lived, evergreen species belonging to the angiosperm family Myrtaceae (Ladiges et al. 2003). They occur naturally from sea level to the alpine tree line, from high rainfall to semi-arid zones and from the tropics to latitudes as high as 43° south (Eldridge et al. 1993). Eucalypts are dominant or co-dominant in almost all vegetation types where they occur (Wiltshire 2004) and are considered keystone species for ecological studies in their natural ranges (Doughty 2000).

Eucalypts are grown as exotic plantation species in tropical and subtropical regions of Africa, South America and Asia, and, where climate allows it, in temperate regions of Europe, South America and Australia. They are utilized for a diverse array of products including sawn timber, mine props, poles, firewood, pulp, charcoal, essential oils, honey and tannin as well as for shade, shelter and soil reclamation. In rural communities of many developing countries, eucalypt wood is an important source of fuel and building material (Eldridge et al. 1993). The expansion of eucalypt plantations throughout the world during the past 60 years can mostly be attributed to the superior fiber and pulping properties of eucalypt species and the increased global demand for short-fiber pulp (Turnbull 1999). In today's "new carbon economy", eucalypts are receiving attention as fast-growing, short-rotation, renewable biomass crops for energy production.

Despite their commercial and ecological importance, genomic resources are only slowly becoming available for eucalypt species. A public genome sequencing effort would facilitate the development and application of such resources and leverage the release of several private resources into the public domain. There are seven major justifications for the generation of a draft sequence of the *Eucalyptus* genome by the US Department of Energy (through the Laboratory Science Program of the Joint Genomes Initiative):

1. ***Eucalyptus* species are desirable for the "Genomes to Energy" sequencing program.** Eucalypts are intensively used to produce energy from charcoal that reduces the net production of greenhouse gases. Eucalypts sequester carbon at an average rate of 10 tons of carbon/hectare/year from planting to harvesting and up to 14 tons/ha/yr in fast-growing tropical plantations (Marcolin et al. 2002). Furthermore, eucalypts have a positive net carbon balance even when computing the production of CO₂ when used for energy from charcoal or as pulp and paper. Eucalypts remove CO₂ from the atmosphere at a rate of 1.8 ton CO₂/ton of dry wood and generate O₂ at a rate of 1.3 ton O₂ per ton of dry wood (Fearnside 1999; Ferreira 2000).
2. ***Eucalyptus* is the most valuable and widely planted hardwood crop in the world because of its superior growth, adaptability, and wood properties.** *Eucalyptus* species occupy 18 million hectares in 90 countries (FAO 2000). Fast growth rates and wide adaptability have contributed to the great interest that *Eucalyptus* receives in many countries outside its native range. Besides their fast growth that allows for shorter rotations, many eucalypt species display wood properties, such as high density, that make them very suitable for fuel and charcoal production, pulp and paper manufacturing, and sawn wood.
3. ***Eucalyptus* biology allows experiments that are difficult or impossible in other forest trees.** *Eucalyptus* species have relatively small genomes (~600 Mbp), no evidence of recent polyploidization and exhibit perfect behavior as diploids with a fully conserved number of chromosomes ($2n = 22$) across all of the planted species. They can be induced to flower in less than 12 months (Griffin et al. 1993; Hasan and Reid 1995) and as early as 14 weeks in species such as *E. occidentalis* (Bolotin et al. 1975). Eucalypts are typically outcrossing, but have hemaphroditic flowers and selfing is possible. *Eucalyptus* trees can be cloned easily in vitro and by the well established mini-cutting technology, thus facilitating replication and extensive testing in different environments. Several commercial eucalypt species have now been genetically transformed, including *E. grandis* and its hybrids (Tournier et al. 2003), *E. camaldulensis* (Mullins et al. 1997; Ho et al. 1998; Valerio et al. 2003) and *E. globulus* (Moralejo et al. 1998), although the rates of success vary greatly among genotypes within species.
4. ***Eucalyptus* species have good basic genomic resources.** Several genetic maps with hundreds of public microsatellites and many known genes have been published. Extensive genome synteny has been observed across species. Many QTLs are known for traits related to production forestry as well as basic adaptability, development and control of flowering. Databases of expressed sequence tags and microarray expression experiments have been produced by private groups and public institutions. Many basic characteristics of the eucalypt genome are known from sample sequencing as well as BAC shotgun sequencing and assembly. Large networks of experimental populations in multiple environments involving segregating progenies and association mapping populations are in place and have been

extensively phenotyped for adaptability and wood quality traits (reviewed in Poke et al. 2005; Myburg et al. 2006). Transformation model systems for *Eucalyptus* have been developed in Japan (Oji Paper), at the University of São Paulo in Brazil and at the University of Melbourne in Australia. These systems will be made available to the research community for functional genomics research. A draft genome sequence is currently being generated for *E. camaldulensis* in a state government funded project at KAZUSA DNA Research Institute in Japan. This genome sequence will almost certainly become public upon, or just before the final release of the proposed DOE sequence, providing opportunities for comparative genomics in *Eucalyptus*, on par with that in rice.

5. ***Eucalyptus* species benefit from a large, diverse and collaborative scientific community worldwide.** Due to its ample genetic variation both at the intra and interspecific levels, its unique evolutionary history, its multiple uses in industry and small communities, *Eucalyptus* trees have received the attention of hundreds of geneticists, breeders, plant physiologists and wood engineers throughout the world. The recently formed *Eucalyptus* Genome Network (www.ieugc.up.ac.za) includes 137 scientists representing 82 research organizations and forestry companies in 18 countries. Although many private companies run active *Eucalyptus* breeding programs, most of the genomic research is conducted by public institutions. In recent years some forest biotech companies have started programs to develop the potential of transgenic technology mainly in lignin modification. There is ample opportunity and existing critical mass to take full advantage of a draft sequence assembly with the derived information to be deposited in the public domain.
6. ***Eucalyptus*, a second tree genome for comparative, perennial plant genomics.** A second tree genome sequence will provide extraordinary opportunities for comparative genomic analysis with the *Populus* genome (and with other non-woody species) to further advance our understanding of the unique facets of tree biology including their perennial growth habit, extensive formation of secondary xylem (wood) and juvenile-mature phase change.
7. ***Eucalyptus* species that could be suitable for solid wood production or for pulp and paper production, have not been found for the continental USA in spite of several large field trials testing diverse genotypes and species.** Growth of *Eucalyptus* in the southeast USA requires resistance to rapidly fluctuating climate changes in late winter and early spring. Genes regulating environmental responses need to be identified and modified so that appropriate genotypes could be produced by a combination of breeding and genetic engineering. Short generation breeding in *Eucalyptus* and a program to produce trees suitable for the USA could be carried out much faster than is typical for forest trees. The efficiency of vegetative propagation in *Eucalyptus* would allow wide-scale planting and rapid deployment of superior clones. If a *Eucalyptus* genomic breeding program were successful, *Eucalyptus* would transform the energy potential of the US, promote economic development in depressed rural areas, rescue the US forest products industry from a long-term decline, and make a substantial and favorable impact on the US carbon budget. It would also be good for the world.

We propose the sequencing of the genome of *Eucalyptus grandis*, the most widely used eucalypt plantation species and an important parent used in the development of fast-growing eucalypt hybrids world-wide. In support of this proposal, we briefly discuss the taxonomic, economic and environmental importance of *Eucalyptus* and propose an action plan for genome sequencing, annotation and dissemination of genome information.

2. Taxonomic importance of *Eucalyptus*

The genus *Eucalyptus* contains in excess of 700 species. The genus belongs to the predominantly southern hemisphere family Myrtaceae, which is embedded in one of the early diverging lineages of the Rosid group of plants (Fig. 1). A later divergence of the Rosids, the Eurosids I & II, include *Arabidopsis*, *Medicago* and *Populus* for which full genomic sequencing is complete or in progress. The Eurosids I lineage, to which *Populus* belongs, includes other well-known trees such as the oaks and birches. Most plants whose genomes are currently being sequenced have an the annual growth habit, except for *Populus*, the only perennial genome that has been sequenced to date. The family to which *Eucalyptus* belongs includes other important tree genera such as *Melaleuca* and *Metrosideros*, as well as several species used for their oils (*Leptospermum* – tea tree) and fruits (*Psidium* – guava), or as spices (*Pimenta* – allspice; *Syzygium* - cloves) and ornamentals (*Callistemon* – bottlebrush). A *Eucalyptus* genome sequence would be the first representative of the Myrtales (Fig. 1), an important sister lineage to the Eurosids and therefore an informative genome for comparative genomics with members of the Rosids (e.g. *Arabidopsis*, *Populus* and *Medicago*) and the Asterids (e.g. tomato).

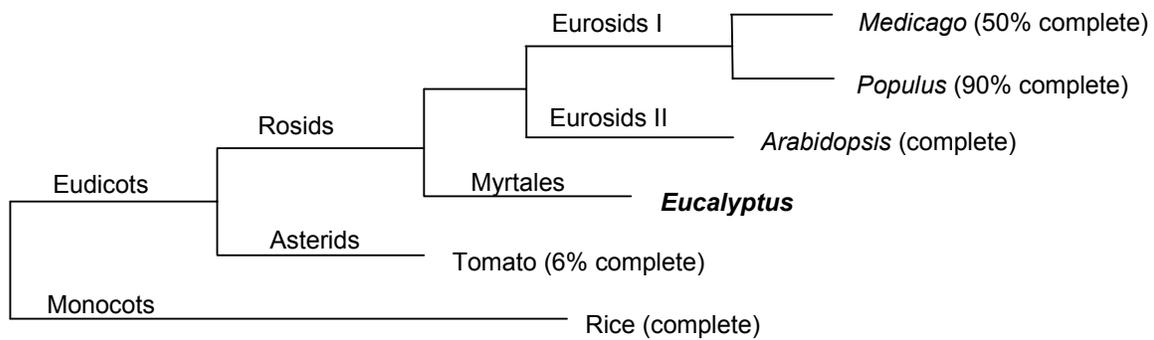


Figure 1. Phylogenetic position (in a simplified cladogram) of *Eucalyptus* relative to other plant species with completed or ongoing genome sequencing efforts.

3. Scientific importance of the *Eucalyptus* genome

There is major scientific interest in *Eucalyptus* in academia and in the private sector. A eucalypt genome sequence will provide significant advances in research methodology for investigations of the ecological and evolutionary biology of eucalypts, as well as in tree breeding. Genomic tools will allow fundamental investigations into the links between gene expression, biochemical pathways and physiological function. Eucalypts are a dominant component of the flora of Australasia and they are keystone ecological species in over 40 million ha of natural eucalypt forests (Eldridge et al. 1993). They have evolved quite separately over millions of years from the dominant tree genera in the tropics and temperate regions of the Northern Hemisphere (for at least 60 million years since the isolation of continental Australia). They grow across a wide range of climatic environments and soil types and are likely to have developed quite unique adaptive strategies to other tree genera such as *Populus*. An important scientific question is whether eucalypts have developed novel solutions for biological challenges associated with being large, long-lived land plants, problems that are shared with other tree genera such as *Populus* and *Pinus*.

Eucalypt genomics will allow investigation of the genetic control of key eco-physiological processes underlying complex responses and multiple interactions between trees and their abiotic and biotic environment. In particular, the presence of eucalypts as keystone species in a range of climatic conditions will allow for investigations into the impact of climate change on ecosystems. A number of eucalypt species are currently showing decline in their natural ranges, perhaps related to increased drought. The eco-physiology of these species in relation to healthy neighboring species is being investigated and this research would be advanced by the availability of comprehensive genomic tools. Other major eco-physiological processes that are particularly tractable in eucalypts include responses to increased salinity, adaptation to low nutrient soils, herbivory, and resistance to insect and pathogen attack. Elucidation of genetic control of ecological processes will enable strategies for sustainable management of ecosystems. Eucalypt genomics will also allow investigation of other processes that are specific to eucalypts, including the production of secondary metabolites that are of commercial value, such as 1,8 cineole (eucalypt oil) and phloroglucinols.

The characteristics of *Eucalyptus* trees that are of greatest scientific importance are their superior fiber qualities, outstanding growth and yield, and excellent carbon sequestration capacity. These factors drive the commercialization of eucalypt plantations in many countries and underlie the environmental benefits of *Eucalyptus* plantation forestry.

4. Economic importance of *Eucalyptus*

In 2000, the FAO estimated a total of 17.9 million hectares of planted *Eucalyptus* worldwide with India the largest planter at over 8 million hectares, followed by Brazil with 3 million hectares (FAO 2000). The majority of these plantations consist of a few eucalypt species and hybrids. Most of the domesticated eucalypts are from the subgenus *Symphomyrtus*, which is the largest of the 10 subgenera currently recognized within *Eucalyptus*, containing over 75% of the species. Four species and their hybrids from this subgenus, *E. grandis*, *E. urophylla*, *E. camaldulensis* and *E. globulus*, account for about 80% of the eucalypt plantations worldwide. Other important species include *E. nitens*, *E. saligna*, *E. deglupta*, *E. pilularis* and *E.*

tereticornis (Eldridge et al. 1993). *E. globulus* is the premier species for temperate zone plantations in Portugal, Spain, Chile and Australia. Elite hybrid clones involving *E. grandis* and *E. urophylla* are extensively used by the pulp and paper industry in tropical and subtropical regions of Brazil, South Africa, Congo and China because of their superior wood quality, rapid growth, canker disease resistance and high volumetric yield.

Although eucalypt breeding is currently a dynamic and technically advanced operation carried out mainly by private companies, eucalypts are still in the early stages of domestication when compared to crop species, with most eucalypt breeding programs only one or two generations removed from the wild. However, due to the ample genetic variation both at the intra and interspecific levels, their short generation times and the ability to clone elite genotypes, eucalypts have quickly become the most advanced genetic material in forestry. Excellent eucalypt breeding programs have been established in countries like Brazil, South Africa, Portugal, Australia and Chile that adopted *Eucalyptus* for industrial plantation forestry.

Productivity values from eucalypt plantations vary by species, site, climate, silviculture and genetics. In subtropical and temperate plantations, mean annual increments (MAI) average between 10 and 20 m³/ha. In the tropics, productivities based on the use of selected clones or seed sources average 30 to 50 m³/ha/yr and reach up to 80 to 100 m³/ha/yr for the best clones on the best sites. In Brazil, the successful combination of hybrid breeding, high selection intensity and clonal forestry, has resulted in a national eucalypt MAI of 38 m³/ha. Even a 10 to 20% increase (e.g. 5 m³/ha/yr) in the global average MAI would produce an additional 90 million cubic meters of eucalypt wood per year. Such increases (and more) will be achievable with a combination of genomics-assisted breeding and genetic engineering of superior eucalypt germplasm.

In an energy economy based on sustained crude oil prices of more than US\$60 per barrel, fast-growing, short-rotation eucalypts would be excellent biomass crops for green energy production. Many of the characteristics of the proposed production of ethanol from plant biomass are remarkably similar to the eucalypt pulp and paper industry, in terms of its focus on cellulose, and its requirements of scale and efficiency. Similar to pulp and paper, such ethanol production facilities will require a continuous supply of material on a year round basis. Annual crops are not suitable for this because harvest and storage can be incompatible with continuous supply. Furthermore, trees favored by the pulp and paper industry can grow on marginal land, not useful for conventional agriculture. A major barrier for the pulp and paper industry, and for biomass fermentation, is the quantity of lignin in wood. Recent advances in the genetic engineering of poplar show that it is possible to increase the ratio of cellulose to lignin by a factor of two and reduce lignin content to as low as 10%. If this modification could be extended to other tree species, commensurate improvements in efficiency could be expected. In addition, a negative correlation of growth and lignin content has been found in studies of poplar, pine and *Eucalyptus*. This correlation indicates that as lignin content decreases, growth rate increases. The availability of genomic research tools will allow scientists to determine the metabolic changes that are required to tailor fast-growing eucalypts for green energy production.

5. Environmental importance of *Eucalyptus*

The "Genomes to Energy" sequencing program is directed to producing energy in ways that reduce the net production of greenhouse gases. *Eucalyptus* species fit very well into this major DOE goal. Eucalypts are very efficient at carbon sequestration with average annual fixation rates of 10 tons of carbon per hectare. Even when considering the CO₂ produced when eucalypts are used for energy in the form of charcoal, eucalypts have a positive net carbon balance. Each eucalypt tree can fix 20 kg of CO₂ per year. This attribute is particularly important when eucalypts are used for reforestation of degraded lands or areas abandoned by intensive agriculture (Marcolin et al. 2002; Fearnside 1999).

Within the Kyoto framework, there are two mechanisms by which planted forests and their wood products can displace the consumption of fossil fuels and make a contribution to reducing net carbon dioxide emissions: (1) Wood products may replace fossil fuels such as mineral coal that traditionally have been used for energy production and (2) wood products may replace materials that require high energy levels to be manufactured. The carbon displaced via this second mechanism is a permanent contribution to CO₂ emission reduction. A comparison was made between CO₂ sequestration and O₂ regeneration when using mineral coal versus eucalypt charcoal in steel production. Mineral coal has a negative impact on both accounts as it generates 1.65 ton of CO₂ and sequesters 1.54 ton of O₂ per ton of steel. In contrast, the eucalypt charcoal sequesters 16.34 tons of CO₂ (10 fold more) and regenerates 11.88 tons of O₂ (8 fold more) from plantation through harvesting to steel production (Ferreira 2000).

Several eucalypt forest based industries in the tropics have implemented long-term projects to supply eucalypt charcoal in substitution to mineral coal for the production of steel, resulting in a significant net reduction of carbon emissions in these regions. For each ton of pig iron made with renewable eucalypt charcoal there is an environmental gain of 3 tons of CO₂ when

compared to fossil fuels. This reduction in emissions is commercialized as carbon credits to be deducted by buyers from their specific emission reduction goals set by the Kyoto protocol (Fearnside 1999). Out of the 4 million hectares currently under *Eucalyptus* plantation in Brazil around 40% is planted by steel industries that convert eucalypt wood into high energy charcoal. In addition to contributing to reduced greenhouse gas emissions, *Eucalyptus* plantations mitigate the pressure on native savanna and rain forest ecosystems traditionally exploited for old-growth forestry. Furthermore, in recent years, a combination of innovative silviculture and management practices for longer 12-year rotation, together with modern drying and saw mill technologies have allowed the production of very high quality eucalypt timber for furniture and construction, thus increasing the contribution of *Eucalyptus* to permanent CO₂ fixation.

Eucalypts are excellent candidates for carbon sequestration and production of energy from marginal land due to their fast growth rate and adaptability to a range of environments including more arid conditions. Availability of land suitable for plantations is a key element of strategies for carbon sequestration and production of bioenergy from trees. Land is available for these activities in the tropics and southern temperate regions. In fact, revegetation and land use change in the form of agroforestry is vital to sustainable natural resource management in many parts of these regions including southern Australia, where amelioration of the effects of changed water hydrology leading to dryland salinity require large scale planting of woody perennials (Australian Greenhouse Office, 2001). Agroforestry provides a large resource for carbon sequestration and production of green energy.

In Australia, estimates of carbon sequestration by eucalypt plantings on average sites are 259 tons CO₂/ha (Harper et al. in review). Carbon trading has seen Japanese companies buy carbon credits from eucalypt plantations in Australia. For example, Tokyo Power Company bought AU\$ 140 million worth of carbon credits from State Forests NSW (O'Neill 1999). Wood gasification techniques have further increased cost-effectiveness and made electricity production from eucalypt biomass attractive (Australian Greenhouse Office). Integrated processing plants (e.g. Narrogin, WA) use modern fluidized bed technology to convert wood to charcoal and then to activated carbon. They also distill eucalyptus oil from the leaves and then use gasification technology on the residues to produce ethanol and steam for energy. The energy produced feeds directly into the electricity grid.

6. Proposed sequencing, annotation and dissemination of genome information

Over the past two years, a series of international meetings of *Eucalyptus* researchers have led to the formation of a genomic research community with more than 100 members (www.ieugc.up.ac.za). In anticipation of a DOE-JGI genome sequencing effort, many more have joined the community, which in future will be known as the *Eucalyptus* Genome Network (EUCAGEN). Several working groups have been established to coordinate different activities of the research community such as genome sequencing, genetic mapping and bioinformatics resource development. The prospect of a second tree genome sequence has also sparked the interest of members of the poplar genome research community, who clearly recognize the value of comparative genomics between poplar and eucalypts. As described in this section, EUCAGEN will collaborate with DOE-JGI in the preparation of biological materials, the development of associated genomic resources such as a fingerprinted BAC library, and the dissemination of genome information.

Motivation for the choice of species (*E. grandis*) to sequence

Eucalyptus grandis is the most widely used species in plantation forestry worldwide in tropical and subtropical areas not only as a pure species, but also as a parental species in hybrid breeding. It displays the fastest growth rates and widest range of adaptability of all other *Eucalyptus* species, most likely due to its extensive natural distribution in Australia. In its natural habitat, *E. grandis* is a medium to very tall tree, 45-55 m high. The southern limit of *E. grandis* is in New South Wales at latitude 32° South. Its distribution then stretches northwards almost continuously in the coastal regions and adjacent ranges until south eastern Queensland (26° South). Further northwards (from approximately 18°S) it occurs in disjunct populations up to the northern extreme of its range on the Windsor Tableland (approximately 16°S). Natural stands typically occur over an altitudinal range from sea level to around 600 m although the more northerly occurrences in Queensland cover an altitudinal range from 400 to 1250 m (Eldridge et al. 1993). The greatest area of plantations of *E. grandis* and its hybrids with other species are those established in Brazil and several other Central and South American countries. It has been planted extensively in India, South Africa, Zambia, Zimbabwe, Tanzania, Uganda and Sri Lanka. Small plantations have been grown in three states of the United States of America, namely California, Florida and Hawaii. Within the subgenus *Symphomyrtus*, *E. grandis* is a member of the section *Latoangulatae*, which also includes other fast-growing subtropical and tropical eucalypts such as *E. saligna* and *E. urophylla*. The International *Eucalyptus* Genome Network (at a business meeting in Pretoria, South

Africa, 9 Nov 2005), selected *E. grandis* for a DOE genome sequencing proposal with *E. globulus* (section Maidenaria) a second choice.

Genome size and structure information

The estimated genome size of *Eucalyptus grandis* is ~600 Mbp (Grattapaglia and Bradshaw 1994). It is a diploid species with a haploid chromosome number of $n = 11$. Due to its preferentially outcrossing mating system and recent domestication, *E. grandis* displays a very high level of nucleotide diversity with approximately one polymorphic site every 70 bp and frequent indels throughout the genome. A sample sequencing study of 3 Mbp of sheared DNA shotgun sequence indicated that on average the *E. grandis* genome has a GC content of 40.15% with gene regions composed of 38.95% introns and 45.51% exons. Of the total bases sequenced, approximately 1.4% were located in transposons, distributed in 310 interspersed repetitive genetic elements, among which 299 were classified as retroelements, mainly LTRs with 244 copia-like and 52 gypsy-like elements. A total of 1636 low complexity sequences and 987 microsatellites including both perfect and interrupted repeats were identified, with an estimated frequency of one microsatellite every 3 kb in the *E. grandis* genome. Only 5.8% of the *Eucalyptus* genome was represented by repetitive elements, lower than the 10% observed in *Arabidopsis*, which is possibly due to the existence of elements not yet described (Lourenço 2004). This low estimate is further supported by an estimate of 4.2% highly repetitive DNA observed in the *E. camaldulensis* genome sequencing effort at KAZUSA DNA Research Institute (based on 1x coverage random shotgun sequencing).

Biological materials to be used for sequencing

Given the very high level of nucleotide diversity in the *Eucalyptus* genome, and an anticipated coverage of 4X, we propose that DNA from a selfed (S_1) tree of *E. grandis* be used for genome sequencing. On average this genome should be 50% homozygous compared to less than 20% in a normal diploid eucalypt. The reduced level of heterozygosity should significantly improve genome assembly of the shotgun reads, as well as the assembly of a physical map of fingerprinted BAC clones to be run in parallel (see below). Although a standard heterozygous genome would provide a snapshot of the level of polymorphism typical of an individual tree, we believe that a better assembly of the 4X draft will be more valuable. Furthermore, a better whole-genome nucleotide diversity estimate could be gained by in silico comparative analysis with the *E. camaldulensis* draft sequence, or by resequencing specific target regions of the genome in a SNP discovery panel. The S_1 *E. grandis* tree to be sequenced will be derived from a breeding program in Brazil and will be made available to any interested party worldwide as in vitro micropropagules. DNA for library construction can be extracted either from leaflet material of in vitro plants, or leaves from mature plants. To minimize potential contamination with organellar genomes, nuclei can be provided to JGI in agarose plugs.

Fingerprinting of BAC clones and construction of a physical map

A public *Eucalyptus grandis* BAC library (5X *Hind*III and 5X *Eco*RI partial digests) with an expected average size of 140 to 160 kbp will be built from the same S_1 genotype in a collaborative effort between the Arizona Genomics Institute (AGI) and GENOLYPTUS, the Brazilian network of genomic research on eucalypts. The expected 55,000 clones will be fluorescently fingerprinted and end sequenced at AGI. AGI has experience in building and fingerprinting of BAC libraries in a high-throughput setting with over 190 BAC libraries constructed to date. It is expected that a high coverage, high quality fingerprint assembly with a minimal number of gaps will be constructed of the S_1 *E. grandis* genotype. The assembled physical map will be published and the BAC library deposited at Arizona Genomics Institute's BAC/EST Resource Center who will supply this library as affordable high-density hybridization filters, clones and libraries to interested parties. With this complementary arrangement it should be possible for any research group in the world to have access to any physical region of the genome for detailed studies. Gaps in the shotgun sequence assembly could be closed by shotgun sequencing of the corresponding regions in the minimum tiling path derived from the physical map.

Sequencing and assembly of the genome sequence

Whole-genome shotgun sequencing will be carried out by JGI. BAC end sequencing will be carried out either by JGI or AGI depending on the final agreements. Genome assembly will be carried out by the bioinformatics group at JGI in collaboration with the Bioinformatics & Evolutionary Genomics Group at Ghent University, the bioinformatics labs at the Catholic University of Brasilia and EMBRAPA in Brazil, the National Bioinformatic Node (NBN) at the University of Pretoria in South Africa, and CSIRO in Australia. JGI will coordinate the effort and distribute the specific tasks.

The Brazilian GENOLYPTUS network will contribute a reference linkage map for *Eucalyptus grandis* with over 400 microsatellite markers derived from genomic and EST sequences mapped on large full-sib pedigrees with over 400 individuals. This map should provide a genome-wide set of anchor loci mapped with high likelihood support for marker order to aid in the assembly of contigs.

The University of Tasmania and University of Pretoria have agreed to collaborate in the construction of a universal genotyping microarray chip for commercially grown eucalypts based on Diversity Array Technology (DART). The DART markers will be mapped separately in large mapping pedigrees ($n > 400$) of *E. grandis* and *E. globulus* in South Africa and Australia, respectively. Approximately 2000 to 3000 markers on the chip are expected to be polymorphic in these pedigrees. The DART fragments corresponding to these markers (single copy in the genome by default) can be sequenced and used to anchor the genetic linkage maps to the genome sequence assembly.

Contribution of ESTs from various sources for genome annotation

Several groups throughout the world will contribute to the annotation effort by providing EST databases and annotation experience. GENOLYPTUS will provide an annotated EST sequence database comprising around 22,000 unigenes derived from sequencing over 20 cDNA libraries from different tissues, species and developmental stages of tree development. CNRS in France will provide another set of over 12,000 differentially expressed ESTs from the temperate species *E. gunnii* and *E. globulus*. The private tree biotech venture Arborgen is currently seeking board approval to put forward unigenes or ESTs from their extensive *Eucalyptus* EST database.

Dissemination of genome information

The draft genome sequence and annotation will be made available through a web browser hosted by JGI. The Bioinformatics & Evolutionary Genomics Group at Ghent University will also develop a web-based bioinformatics resource in collaboration with bioinformatics groups in Brazil, South Africa, and Australia. We anticipate that KAZUSA DNA Research Institute will take part in this effort and that the *E. camaldulensis* genome sequence will be made available for comparative genomics analysis and genome mining in a joint genome browser.

Letters of Support

This proposal is supported by 61 letters from representatives of public and private organizations involved in *Eucalyptus* research (please see Letters of Support attached).

7. Literature Cited

- Australia Greenhouse Office and the Murray Darling Basin (2001) The contribution of mid to low rainfall forestry and agroforestry to greenhouse and National Resource Management outcomes. Australian Greenhouse Office
- Doughty RW (2000) *Eucalyptus*. A natural and commercial history of the gum tree. The Johns Hopkins University Press, Baltimore and London
- Eldridge K, Davidson J, Harwood C, van Wyk G (1993) *Eucalypt* domestication and breeding. Clarendon Press, Oxford
- FAO (2000) Global forest resources assessment 2000 - Main report: FAO Forestry paper <http://www.fao.org/forestry/fo/fra/main/index.jsp>.
- Fearnside PM (1999) Forests and global warming mitigation in Brazil: Opportunities in the Brazilian forest sector for responses to global warming under the "Clean Development Mechanism." *Biomass and Bioenergy* 16(3): 171-189
- Ferreira, O. C. 2000 The future of charcoal in metallurgy - emission of greenhouse effect gases in the production and use of charcoal in metallurgy. *Economy and Energy* 21 (available at http://ecen.com/para_imprimir/eee20.pdf)
- Grattapaglia D, Bradshaw HD (1994) Nuclear DNA content of commercially important *Eucalyptus* species and hybrids. *Canadian Journal of Forest Research* 24:1074-1078
- Griffin AR, Whiteman P, Rudge T, Burgess IP, Moncur M (1993) Effect of paclobutrazol on flower-bud production and vegetative growth in two species of *Eucalyptus*. *Can J For Res* 23: 640-647
- Harper et al. in review. The potential of greenhouse sinks to underwrite improved land management. *Ecological Engineering*.
- Hasan O, Reid JB (1995) Reduction in generation time in *Eucalyptus globulus*. *Plant Growth Regul* 17: 53-60
- Ho CK, Chang SH, Tsay JY, Tsai CJ, Chiang VL, Chen ZZ (1998) *Agrobacterium tumefaciens*-mediated transformation of *Eucalyptus camaldulensis* and production of transgenic plants. *Plant Cell Reports* 17:675-680
- Ladiges PY, Udovicic F, Nelson G (2003) Australian biogeographical connections and the phylogeny of large genera in the plant family Myrtaceae. *J Biogeogr* 30:989-998
- Lourenço (2004) Genomic structure of three mega basepairs of genomic shotgun DNA of *Eucalyptus grandis*: nucleotide content, repetitive sequences and genes. M.Sc. Thesis, Univ of Campinas, Brazil, 117 pp (in Portuguese available at http://www.universiabrasil.net/images/docs/tese_rodrigo.pdf)
- Marcolin M, Santos AA, Salati E (2002) Emissões e remoções de dióxido de carbono por mudanças nos estoques de florestas plantadas. Brazilian Foundation of Sustainable Development, Published by the Brazilian Ministry of Science and Technology. 47 pp. (available at <http://www.mct.gov.br/index.php/content/view/21456.html>)
- Moralejo M, Rochange F, Boudet AM, Teulieres C (1998) Generation of transgenic *Eucalyptus globulus* plantlets through *Agrobacterium tumefaciens* mediated transformation. *Australian Journal of Plant Physiology* 25:207-212
- Mullins KV, Llewellyn DJ, Hartney VJ, Strauss SH, Dennis ES (1997) Regeneration and transformation of *Eucalyptus camaldulensis*. *Plant Cell Reports* 16:787-791
- Myburg AA, Potts B, Marques CM, Kirst M, Gion JM, Grattapaglia D, Grima-Pettenati J (2006) *Eucalyptus*. In: Kole C (ed) *Genome Mapping & Molecular Breeding in Plants Volume 7: Forest Trees*. Springer, Heidelberg, Berlin, New York, Tokyo, pp in press
- Poke FS, Vaillancourt RE, Potts BM, Reid JB (2005) Genomic research in *Eucalyptus*. *Genetica* 125:79-101
- Tournier V, Grat S, Marque C, El Kayal W, Penchel R, de Andrade G, Boudet AM, Teulieres C (2003) An efficient procedure to stably introduce genes into an economically important pulp tree (*Eucalyptus grandis* x *Eucalyptus urophylla*). *Transgenic Research* 12:403-411
- Turnbull JW (1999) *Eucalypt* plantations. *New Forests* 17:37-52
- Valerio L, Carter D, Rodrigues JC, Tournier V, Gominho J, Marque C, Boudet AM, Maunders M, Pereira H, Teulieres C (2003) Down-regulation of cinnamyl alcohol dehydrogenase, a lignification enzyme, in *Eucalyptus camaldulensis*. *Molecular Breeding* 12:157-167
- Wiltshire RJE (2004) *Eucalypts*. In: Burley J, Evans J, Youngquist JA (eds) *Encyclopedia of Forest Science*. Elsevier Science, Oxford, pp 1687-1699