Perspectives on genome mapping and marker assisted breeding of eucalypts

Dario Grattapaglia
- EMBRAPA Recursos Genéticos e Biotecnologia
- Genomics Sciences Program - Universidade Católica de Brasília
1990 - 1996 markers and maps with RAPD

1996 - 1997 AFLP

1997 - 2002 microsatellites

2002 - 2007

New microsatellites (BACends, EST) Fluorescent detection

2007 SFPs in oligoarrays and DArT
Clonal identification with *Eucalyptus* microsatellites
Seed orchard

Commercial plantation by half-sib family blocks or half-sib progeny trial

Identification of top trees for specific phenotypic traits of interest

Paternity testing of selected trees using microsatellite markers to precisely identify their pollen parents

Retrospective selection of parents with high specific combining ability (SCA) to be used in controlled crosses and/or to cull from seed orchard parents of low SCA
The tree is the real pulp mill

- Increased forest productivity: $m^3$ wood/hectare/year
- Reduction in wood specific consumption for pulp production: $m^3$ of wood/ton of cellulose pulp
- Combination of large wood volume per hectare with low specific consumption of wood per ton of pulp: reduced industrial costs

Breeding is key to competitiveness

Genomics will be one more tool available to the breeder
Eucalyptus species display wide genetic variation for chemical and physical wood traits.

Contrasting species are genetically and genomically very close, can be easily interbred and generate fertile intermediate hybrids.

What are the genetic and molecular bases of the observed phenotypic differences?

- Differences in coding regions: enzymes?
- Differences in regulatory elements: cis, trans, microRNA?

By understanding these differences it should be possible to exploit the natural variation in a more directed way through molecular/conventional breeding.
Genomic variation between individuals is larger than previously thought and much greater in non-coding regions.

- Basic assumptions in genetics: genomes of individuals of the same species are collinear at the sequence level and contain the same genic complement.
- Exceptions include SNPs, indels, translocations, and transposons insertions.
- This assumption has been challenged in recent years as we have been able to sequence and compare long stretches of DNA or even complete genomes.
- Role of the repetitive, non-coding portion of the genomes (junk DNA) in the regulation of gene expression and impact on phenotypic variation.
- Consistent with phenotypic variation dependent on the position effect of transgene insertion.
Sequence sharing at 4 loci (2.8 Mbp of sequence) between two maize inbreds

On average ~50% of the compared sequences are NOT collinear mainly due to retrotransposons insertions. Differences in repetitive DNA genome content and distribution supply different genome environments that in turn affect specificity and temporal regulation of gene expression. These differences have been proposed as the explanation to heterotic complementation in maize.
Genomic variation among individuals, junk DNA and pan-genome

- Sequencing a single genome is interesting: sequencing and comparing several genomes is much more
- Comparative large scale genomics and analysis of microcollinearity will be an increasingly interesting area to be explored to understand and manipulate phenotypic variation
- Concept of the PAN-GENOME: a core shared portion and a dispensable individual specific portion will be needed to characterize a genome
- The developing view of transcriptional regulation as a complex and modular system, in which long-range interactions and the involvement of transposable elements are frequently observed, lends support to the possibility of an important functional role for the dispensable genome (Morgante et al. 2007)
- Important role of forward genomics by QTL mapping: let the phenotype tell us where the key genes are!
Large scale phenotyping in segregating populations

Genetic map of transferable markers and QTLs

High resolution genetic map (SNPs)

Minimum tiling path physical map (BAC Fingerprinting)

Identification and validation of target genes

Cross validated QTL LOD >> 3,0

~ 10 cM

< 1.0 cM

~ 500 a 1000 kpb

Genetic mapping of candidate genes

Association mapping (whole genome scan, candidate genes)

Full genome sequence assembly

Functional tests by transformation (complementation, superexpression, interference)

Forward genomics approach from phenotypes to genes
Participating institutions in the GENOLYPTUS project
(Brazilian Network of Eucalyptus Genome Research)
Mating design used to generate segregating families for QTL detection and validation and family based association mapping.

Crosses made with elite parents from different companies.

<table>
<thead>
<tr>
<th><strong>E. grandis</strong> Coffs Harb. VCP</th>
<th><strong>E. urophylla</strong> (Flores) IP</th>
<th><strong>E. globulus</strong> K-Riocell</th>
<th><strong>E. dunnii</strong> Rigesa</th>
<th><strong>E. camaldulensis</strong> V-Mannesmann</th>
<th><strong>E. uro. x E. glob. K-Riocell</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>G1 x G2 (est. VCP x pol. AR)</td>
<td>U1 x G2 (est. IP x pol. AR)</td>
<td>G2 x GL1 (est. AR x pol. K-R)</td>
<td>D1 x G2 (est. RG x pol. AR)</td>
<td>G2 x C1 (est. AR x pol. V-M)</td>
<td>(GxD) x G2 (est. K-R x pol. K-R)</td>
</tr>
<tr>
<td>G1 x U2 (est. VCP x pol. CE)</td>
<td>U1 x U2 (est. IP x pol. CE)</td>
<td>U1 x GL2 (est. IP x pol. K-R)</td>
<td>D1 x U2 (est. RG x pol. CE)</td>
<td>C1 x C1 (est. CE x pol. K-R)</td>
<td>(GxD) x U2 (est. K-R x pol. CE)</td>
</tr>
<tr>
<td>G1 x D2 (est. VCP x pol. K-R)</td>
<td>U1 x D2 (est. IP x pol. K-R)</td>
<td>D1 x D2 (est. RG x pol. K-R)</td>
<td>D1 x (UxGL)</td>
<td>U1 x GL2 (est. IP x pol. V-M)</td>
<td>(GxD) x D2 (est. K-R x pol. K-R)</td>
</tr>
</tbody>
</table>

Mating design used to generate segregating families for QTL detection and validation and family based association mapping.
Main experiments

Complementary experiments

- Planted 22 segregating families for QTL mapping and validation
- Over 1000 trees per family for 13 families
- Several hybrid combinations involving *E. globulus*
E. grandis x E. dunnii X E. urophylla x E. globulus hybrid

São Paulo experiment

Rio Grande dos Sul experiment
E. grandis x E. dunnii X E. camaldulensis hybrid
Age 36 months
Sample collection for phenotypic analysis
~35,000 trees sampled in 2006

COPYRIGHT DARIO GRATTAPAGLIA
Calibration curves for some main traits

Total lignin

Syringyl/ Guaiacyl lignin ratio

Pulp yield

Leonardo Chagas de Souza, UFV
Comparative QTL mapping for wood properties

Data from three unrelated families

Currently these QTL targets are being validated in five further unrelated families involving several different parents and species.
Application of QTL information for marker assisted selection

Hybrid vigor at the family level

Hybrid vigor at the individual level
Mating between plus hybrid trees to maximize segregation for several traits in the outbred F2

QTL MAPPING STAGE

QTL mapping for wood properties traits such as lignin, fiber and wood density

Genotyping > 1000 seedlings with a small (~6 to 10) set of flanking markers for a targeted number of QTL for wood quality traits

MAS STAGE

Deployment of a large number (> 1000) of F2 progeny individuals to maximize probability of generating a recombinant individual with a superior multiple QTL allele content

Early marker assisted selection. Selection intensity is increased by MAS for late expressing traits but number of trees commonly deployed in progeny (~100) test is kept the same, thus allowing large variation to select for other traits such as volume growth, form and branching habit.
The extent of LD in *Eucalyptus* is very limited

- Very high nucleotide diversity: 1 SNP every ~70 bp
- Polymorphic SNPs between and within species (also fixed within species)
- Significant intragenic LD does not extend over 200 bp in a 1.7 kb CCR segment and a 2.7 kb CAD segment - no haplotypic blocks seen
- The number of TagSNPs needed to capture the whole variation remains very high
- Whole genome scan with SNPs likely to be impractical with current funding - 600,000 SNPs needed in best scenario
- Strategic importance of positional candidate gene discovery by QTL mapping, e-QTL analysis together with gene mapping
Association mapping approaches

✓ WHOLE GENOME SCAN
  ✓ How many markers do you need?

✓ CANDIDATE GENE BASED
  ✓ How good are the candidate genes?
Sequences by species

- *E. grandis* 51%
- *E. globulus* 16%
- *E. urophylla* 9%
- *E. pellita* 10%
- *Eucalyptus sp.* 14%
- *Eucalyptus sp.* 14%
- *Eucalyptus sp.* 14%
Table 1. Overview of the EST data set

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Code</th>
<th><em>Populus</em> genotype</th>
<th>ESTs, n</th>
<th>Average length,* bp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cambial zone</td>
<td>A + B</td>
<td><em>tremula × tremuloides</em></td>
<td>5,326</td>
<td>367</td>
</tr>
<tr>
<td>Active cambium</td>
<td>UB</td>
<td><em>tremula</em></td>
<td>4,647</td>
<td>366</td>
</tr>
<tr>
<td>Dormant cambium</td>
<td>UA</td>
<td><em>tremula</em></td>
<td>3,655</td>
<td>403</td>
</tr>
<tr>
<td>Tension wood</td>
<td>G</td>
<td><em>tremula × tremuloides</em></td>
<td>5,723</td>
<td>408</td>
</tr>
<tr>
<td>Wood cell death</td>
<td>X</td>
<td><em>tremula × tremuloides</em></td>
<td>4,867</td>
<td>548</td>
</tr>
<tr>
<td>Young leaves</td>
<td>C</td>
<td><em>tremula × tremuloides</em></td>
<td>5,013</td>
<td>351</td>
</tr>
<tr>
<td>Senescing leaves</td>
<td>I</td>
<td><em>tremula</em></td>
<td>5,726</td>
<td>366</td>
</tr>
<tr>
<td>Cold-stressed leaves</td>
<td>L</td>
<td><em>tremula × tremuloides</em></td>
<td>4,066</td>
<td>448</td>
</tr>
<tr>
<td>Dormant buds</td>
<td>Q</td>
<td><em>tremula</em></td>
<td>5,815</td>
<td>558</td>
</tr>
<tr>
<td>Petioles</td>
<td>P</td>
<td><em>tremula</em></td>
<td>6,443</td>
<td>559</td>
</tr>
<tr>
<td>Virus/fungal-infected leaves</td>
<td>Y</td>
<td><em>tremula</em></td>
<td>1,395</td>
<td>438</td>
</tr>
<tr>
<td>Floral buds</td>
<td>F</td>
<td><em>trichocarpa</em></td>
<td>6,760</td>
<td>351</td>
</tr>
<tr>
<td>Female catkins</td>
<td>M</td>
<td><em>trichocarpa</em></td>
<td>6,112</td>
<td>553</td>
</tr>
<tr>
<td>Male catkins</td>
<td>V</td>
<td><em>trichocarpa</em></td>
<td>4,855</td>
<td>485</td>
</tr>
<tr>
<td>Apical shoot</td>
<td>K</td>
<td><em>tremula × tremuloides</em></td>
<td>5,380</td>
<td>481</td>
</tr>
<tr>
<td>Shoot meristem</td>
<td>T</td>
<td><em>tremula × tremuloides</em></td>
<td>8,371</td>
<td>535</td>
</tr>
<tr>
<td>Bark</td>
<td>N</td>
<td><em>tremula × tremuloides</em></td>
<td>4,891</td>
<td>548</td>
</tr>
<tr>
<td>Roots</td>
<td>R</td>
<td><em>tremula × tremuloides</em></td>
<td>5,786</td>
<td>593</td>
</tr>
<tr>
<td>Imbied seeds</td>
<td>S</td>
<td><em>tremula</em></td>
<td>6,188</td>
<td>502</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>Normalized</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>25,218</td>
<td></td>
<td>1,000</td>
<td>6,784</td>
<td>466±83</td>
</tr>
<tr>
<td>55,531</td>
<td></td>
<td>8,075</td>
<td>459±56</td>
<td></td>
</tr>
</tbody>
</table>

*The part of the sequences that passed the PARACEL TRANSCRIPTASSEMBLER filters. The average read length has increased substantially during the course of this work.*
Screening of 3 new EST microsatellites in a panel of 8 trees involving 6 *Eucalyptus* species
Sequencing of approximately 140 million bases
Obtained > 1,000,000 valid reads with over 100 bases
In a total of a week ~30,000 US$
Candidate gene selection for association mapping experiments

 ✓ Knowledge of function from biochemical role

 ✓ Inference of function based on similarity measures with other sequences in Genebank

 ✓ Transformation experiments: interference, knock-outs, activation tagging

 ✓ Co-localization of genes with QTLs mapped for traits using draft assembly
  taking into consideration the fact that the probability of the gene being a candidate depends on the precision with which the QTL is mapped

 ✓ Differential expression of the gene between segregating contrasting phenotypes in e-QTL experiments

 ... or do all the genes if some form of SNP genotyping platform becomes affordable!!
Difference between tissues

Differences between genotypes

Differences between species

E. grandis

E. globulus

E. grandis

E. globulus

GENOLYPTUS Chip
Pilot experiment
### GENOLYPTUS Chip pilot experiment

<table>
<thead>
<tr>
<th>Eucalyptus grandis (GR)</th>
<th>Eucalyptus globulus (GL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Fast growth</td>
<td>- Slow growth</td>
</tr>
<tr>
<td>- High adaptability</td>
<td>- Low adaptability in the tropics</td>
</tr>
<tr>
<td>- High lignin</td>
<td>- Low lignin</td>
</tr>
<tr>
<td>- Low cellulose yield</td>
<td>- High cellulose yield</td>
</tr>
</tbody>
</table>

**A1**  
**A2**  
**B1**  
**B2**  

---

**COPYRIGHT DARIO GRATTAPAGLIA**
~ 800 bp EST contig

- Each probe synthesized twice on chip
- Randomized distribution of probes onto the chip
Variation between different genotypes of *E. grandis* xylem
Interspecific variation: *E. globulus* x *E. grandis*

**TIR-NBS** disease resistance-like protein [(Populus tomentosa x P. brefortiana) x P. trichocarpa]---contains similarity to SCUTL putative aquaporin TFPI: Vitis berlandieri x Vitis rupestris.


**SKP1 component** [Medicago truncatula]


Similar to gb|X84260| POS5 gene product from Saccharomyces cerevisiae. EST gb|W43875| zinc transporter [Eucalyptus grandis]
Segregation of SNP haplotypes in differentially expressed genes allow genetic mapping them.
Co-location of differentially expressed genes (E. grandis x E. globulus) with QTLs

Lignin

LOD 4,001
Expression QTL analysis of S-adenosylmethionine synthase (SAMS, black line) and adenosylhomocysteinase (SAH, grey line) in an E. grandis × E. globulus. The genetic map positions of the genes (black triangles) suggest that SAMS is cis-regulated, while SAH is controlled in trans by three loci (gray lines peaks of LR~5) located in different genomic regions. (Kirst et al. 2005)
Press Release: June 8, 2007

DOE Joint Genome Institute Announces 2008 Genome Sequencing Targets

Eucalyptus, Foxtail Millet, Red Algae, and Novel Microbial Communities Added to Growing Bioenergy and Carbon Cycling Portfolio

WALNUT CREEK, CA—Toward the goal of harnessing the power of nature through DNA sequencing, the DOE Joint Genome Institute (DOE JGI) has announced the latest Community Sequencing Program (CSP) portfolio. These plant and microbial targets—most with implications for helping reduce the nation’s dependence on fossil fuel—total some 21 billion nucleotides of DNA sequence capacity allocated to public projects submitted through the CSP for fiscal year 2008.

“This year’s selections are completely aligned with the CSP mission, that is, selecting DOE-relevant organisms with the largest and diverse communities of investigators,” said Jim Brisset, DOE JGI Deputy Director and manager of the CSP. “The response to this year’s program, with over 120 submissions, demonstrates an increasing desire to fuel discovery with DNA sequence information—which DOE JGI makes freely available through its web portals and the public databases.”

Among the highest profile of these projects, and largest, with a 600-million-nucleotide genome, is the eucalyptus tree genome—geared to the generation of resources for renewable energy—led by Alexander Myburg of the University of Pretoria, South Africa, with Gerald Tuskan of Oak Ridge National Laboratory (and DOE JGI), and Darío Grattapaglia, of EMBRAPA Genetic Resources and Biotechnology (Brazil).
Perspectives of applied genomics in *Eucalyptus* breeding

- Genomic resources will be abundant and public

- Improvements and cost reduction of genomic methods: it will be feasible to resequence several genomes for association genetics

- Complete reference sequence of *E. grandis* Brasuz1 will be very valuable

- Main limitation: availability of appropriate structured material, sufficiently replicated across field sites and precisely phenotyped for traits of interest

- Research institutions and industry with such materials will be in best position to collect the highest gains with the technology

- Given the extraordinary genetic variation in Eucalyptus, the ingenuity of breeders and powerful genomic tools the perspectives of applied genomics in Eucalyptus are encouraging
Some thoughts on the use of GM Eucalyptus

- **Technical issues**
  - Targets of modification: what genes and how to modify?
  - Transformation protocols: genotype dependence, efficiency
  - Long term stability and tissue specificity of gene expression
  - Tree sterility, gene flow, propagation system

- **Economic issues**
  - Relative gain: GM x natural variation
  - Intellectual property: patent licensing, freedom to operate, royalties
  - Biosafety: GM trial time + rotation time (harvest) x duration of patent
  - Who has the technology x who has the genetic material
  - Investment risk: regulation and biosafety issues still unclear

- **Political issues, public perception, client demand**
  - Forest certification (FSC)

However, GM trees like in annual crops could represent a technology divide and necessary condition for a forest based industry to remain competitive.
Vision for 2010: GENOLYPTUS
an integrated database of genomic and phenotypic data

Network of experimental populations with continued measurements and quantitative genetics analyses

Multiple-pedigree validated QTLs; anchored genetic maps and physical contigs

Candidate genes, QTLs, microsatellite markers positioned onto genome sequence

Variability in genes within and between Eucalyptus species
Acknowledgments

This is the work of a lot of people!

- All university and company scientists, graduate and undergraduate students that participate in the GENOLYPTUS project

- All 22 institutions that constitute the GENOLYPTUS project

- Brazilian Ministry of Science and Technology
  - FINEP – Fundo Verde Amarelo
  - CNPq