What is this presentation about?

Features of the *Eucalyptus* genome that impact genomic, genetics and breeding research

**Eucalyptus genome projects**

- *E. camaldulensis* (Exsertaria), 2012, 654Mb, assembly
- *E. grandis* (Latoangulatae), 2014, 691Mb, draft
- *E. globulus* (Maidenaria), unpublished, 535Mb

**Genomic features impacting genetic research**

- **Meiotic recombination**
  - Knowledge about recombination in Eucalyptus comes from maps
  - Nothing is known about how recombination varies across the genome among individuals within a population in Eucalyptus

- **Linkage disequilibrium**
  - Knowledge from short genic segments showing a rapid decay after 500-1500 bp
  - No attempt made to use LD to describe past demographic events and the life history of the eucalypts

- **Nucleotide diversity**
  - Described within a limited set of genes or pooled cDNA revealing generally high but variable rates (0.006 up to 0.05)

**Opportunities in genomic research in Eucalyptus**

- Very high synteny and colinearity among *Eucalyptus* genomes
- Genome-wide genotyping platforms already developed
  - DArT Array - 7,680 probes
  - Infinium EuCHIP60k - 60,090 SNPs
  - RAPID Genomics Sequence capture – 40,000 probes
- DNA Sequencing is accessible at feasible prices

**Genome-wide patterns of recombination, linkage disequilibrium and nucleotide diversity from pooled resequencing and single nucleotide polymorphism genotyping unlock the evolutionary history of Eucalyptus grandis**

Orzenil Bonfim Silva-Jr. and Dario Grattapaglia

### Aim
- Examine the recombination rate, nucleotide diversity, extent of LD and rate of mutation in the genome
- Infer past demographic fluctuations and reconstruct the life history of the species

### Plant material
- Sample of 48 unrelated *Eucalyptus grandis* trees, from two wild populations in Australia (Atherton) and Coffs Harbor.
- Linkage mapping: 189 F1 individuals (*E. grandis x E. urophylla*)

### Data
- Whole genome pooled resequencing: 89 Gb filtered (100 bp PE)
- Genome-wide Infinium SNP Genotyping: 21,351 SNPs (MAF>0.05)

### Methods

#### General workflow
- N = 36
  - n = 72
  - ρ
  - θ_w
- Pooled sequencing
- N = 189 F1 *E. grandis* cross
  - n = 378
  - c
  - μ
  - r^2
  - ρ
- N = 48
  - n = 96

#### Pooled sequencing data from 36 trees of *Eucalyptus grandis*

<table>
<thead>
<tr>
<th>Sample</th>
<th>Subgenus/Section</th>
<th>Species/Source</th>
<th>N</th>
<th>n</th>
<th>Passed filters reads</th>
<th>High quality aligned reads</th>
<th>High quality aligned bases</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Symphyomyrtus/Latoangulatae</td>
<td><em>E. grandis</em> – IP</td>
<td>425,411,478</td>
<td>326,001,913</td>
<td>27,316,962,328</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Symphyomyrtus/Latoangulatae</td>
<td><em>E. grandis</em> – CH</td>
<td>532,480,914</td>
<td>377,576,654</td>
<td>31,729,219,343</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Symphyomyrtus/Latoangulatae</td>
<td><em>E. grandis</em> – AT</td>
<td>499,189,920</td>
<td>385,799,954</td>
<td>32,370,520,523</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Each pool was sequenced at > 70x depth
Each individual to an average 6x depth

#### SNP frequency spectrum across 13M sites from resequencing and 36k sites from genotyping

*Multi-species SNP selection mitigated ascertainment bias*

- EuHIP60K SNPs
- All 13 M SNPs in *E. grandis*
- Average of 1,000 random SNP sets of same size

No significant difference was seen among the three site frequency spectra (KS nonparametric test)

#### Pooled Sequence data analysis

**Step 1: effective depth of coverage**

- Reasonable total depth of coverage (> 6 per sample level (pool))
- Equal contribution of each pool to the read depth
- Reasonable total depth of coverage (>30) summed over all samples (pools)
- Exclude regions of bad read mappability

---

#### Linkage mapping
- Quantified c (crossover/bp/generation) from high density maps
- Recombination rate
- Explored the variability of ρ at variable inter-SNP distances
- Quantified average ρ at relevant genomic scales
- Identified and GO annotated recombination hotspots
- Nucleotide diversity
- Quantified the average mutation rate from θ = 4Neµ (sequencing)

#### Linkage disequilibrium
- Decay of LD at the genome-wide scale (21 million pairwise distances)
- LD decay at variable MAF, corrected for structure and relatedness

#### Life history
- Under the neutral model, quantified the variation of Ne
- Estimated TMRCA for *E. grandis* from ARGs built for large haplotypes
**Pooled Sequence data analysis**

*Step 2: resolution of the SNP frequency spectrum*

\[
\hat{\pi}_{SNPs} (x) = \frac{\sum (x^2 - 1)}{\sum (x - 1)}
\]

Accept heterozygous sites at PALT>90% according to APS

Allele-frequency estimation accounts for site-specific error rates and sampling bias

---

**Genotyping data analysis and Haplotype Phasing**

Two independent high density SNP maps (8,387 SNPs) revealed assembly inconsistencies in the current version of the *Eucalyptus grandis* genome (corrected in V 2.0)

---

**High-density linkage maps aligned to the *E. grandis* reference genome**

*E. grandis* 4,396 SNPs, 1,310.2 cM

*E. urophylla* 3,991 SNPs; 1,352.8 cM

*Eucalyptus grandis* reference genome - 603.8 Mb covered (99.8%) and \( \epsilon = 3.18 \times 10^8 \text{ bp}^2 \text{ generation}^{-1} \) (2.62 for the updated genome V 2.0)

---

**Genome-wide estimates of population-scaled recombination rate (\( \rho \)) using different SNP data sets and estimation methods for *E. grandis***

<table>
<thead>
<tr>
<th>Data set type</th>
<th># SNPs surveyed</th>
<th>Estimation method</th>
<th>( \rho ) (x 10^-3 bp^-1)</th>
<th>s.d. (x 10^-3 bp^-1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infinium SNPs</td>
<td>21,351</td>
<td>Hotspotter</td>
<td>1.125</td>
<td>1.69</td>
</tr>
<tr>
<td>Infinium SNPs</td>
<td>21,351</td>
<td>LDHAT</td>
<td>0.530</td>
<td>0.460</td>
</tr>
<tr>
<td>Infinium SNPs</td>
<td>21,351</td>
<td>LD - Hill &amp; Weir</td>
<td>0.690</td>
<td>0.156</td>
</tr>
<tr>
<td>Pooled Sequencing SNPs</td>
<td>13,000,000</td>
<td>mlRho</td>
<td>1.470</td>
<td>10.0</td>
</tr>
</tbody>
</table>

Estimated recombination rate based on more distal markers agree between theoretical models independently of the number of segregating sites

---

**Estimates of \( \theta_e \) effective population sizes (N_e), nucleotide diversity (\( \pi_w \)), mutation rate per generation (\( \mu \)) based on the whole-genome pooled resequencing data**

<table>
<thead>
<tr>
<th>SCALE (inter-SNP distance)</th>
<th>( \rho ) (x 10^-3)</th>
<th>( \pi_w ) (x 10^-6)</th>
<th>( \mu ) gen. (x 10^-9)</th>
<th>( \mu / \epsilon )</th>
<th>( \theta_e ) mlRho bp^-1</th>
<th>( \theta_e ) Pooool. bp^-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 – 100 bp</td>
<td>142.1</td>
<td>1,116,352</td>
<td>0.496</td>
<td>0.16</td>
<td>0.024</td>
<td>0.022</td>
</tr>
<tr>
<td>0 – 2 kb</td>
<td>14.3</td>
<td>122,421</td>
<td>4.93</td>
<td>1.55</td>
<td>0.040</td>
<td>0.040</td>
</tr>
<tr>
<td>0 – 50 kb</td>
<td>1.47</td>
<td>11,557</td>
<td>47.94</td>
<td>15.98</td>
<td>0.060</td>
<td>0.060</td>
</tr>
</tbody>
</table>

Estimated recombination rate between close markers (less than 2 kb apart) is 10 times higher than predicted by theoretical models based on more distal markers
Main findings: genomic features and life history

- Genomic features ($\rho$, $\theta_w$, $\mu$, LD, $N_e$) were estimated genome-wide using millions of SNPs by whole-genome sequencing and genotyping.
- At the genome-wide scale, mutation is more important than recombination ($\rho/\theta_w = 0.65$) in shaping the genetic variation.
- The variation in $N_e$ as a consequence of varying recombination across genomic distances inform about demographic history:
  - Using $N_e = 112,421$ the TMRCA from ARGs was dated at 3.2 MYA (1.7-4.8 MYA) consistent with fossil records that place E. grandis lineage divergence at 2.5 MYA (Ladiges et al., 2003).
  - Using $N_e = 11,557$ the TMRCA from ARGs was dated at 469.5 KYA (434.7-504.2 KYA) consistent with the geological dating of the first age in the southern hemisphere (390-730 KYA).

Main findings: impacts on applied genomics

- The extent of LD decay assessed at the genome-wide scale is slower than previously reported from candidate gene studies.
- Genome-wide usable LD within ~4-6 kb: ~10¹ SNPs should provide satisfactory marker density for GWAS studies.
- However the slower LD decay should complicate the precise pinpointing of causative polymorphisms (QTN).
- In breeding populations (up to effective sizes $N_e \approx 100$) with more extensive LD, low density genotyping system (3 – 5k SNPs) should provide abundant power to capture the majority of linked effects in genomic predictive models.
Thank you!

orzenil.silva@embrapa.br