Genome Target Sequencing in Loblolly Pine (*Pinus taeda* L.) using Different Multiplexing Strategies

Mengmeng Lu[1], Carol Loopstra[1], Kostantin Krutovsky[1,2]
[1]Texas A&M University, College Station, TX;
[2] Georg-August-University of Göttingen, Göttingen, Germany

Objective of this research

1. To develop efficient strategy to genotype multiple individual trees.

2. To test the capture efficiency under non-multiplexed and multiplexed conditions.

3. To find optimal multiplexing.

Methods

We applied the Agilent SureSelect Target Enrichment method to capture unigene-based targeted genomic sequences in loblolly pine (*Pinus taeda* L). We used 35,386 unigenes that were assembled by Dr. Chun Liang (Miami University, Oxford, Ohio) to design 647,634 oligonucleotide baits. Two single (A and B) and two multiplexed (C and D) DNA libraries were constructed and sequenced to test two multiplexing strategies: A and B were non-multiplexed samples representing DNA of the megagametophyte and embryo from a single seed, respectively, while C and D were multiplexed pools composed of four and eight indexed individual DNA samples of megagametophytes, embryos or needles from two and seven individual trees, respectively. Each library was hybridized to the same number of baits. After capturing the targeted sequences, all samples were sequenced in Illumina Hiseq2000 using paired-end sequencing (2×100 bp). A and B were pooled and sequenced in lane 1, while C and D were sequenced in lanes 2 and 3, respectively.

Results

1. 70M, 275M and 234M reads (one direction) were outputted from lanes 1, 2 and 3, respectively. ~13% of the reads from lane 1 were discarded after FastQC check, while ~27% from each of lanes 2 and 3 were discarded.

2. High quality reads were mapped to the original unigenes and to the draft loblolly pine reference genome assembly (v0.9, provided by the PineRefSeq project) using BWA and SAMtools. Mapping percentage was calculated by BAMtools (Fig. 1).

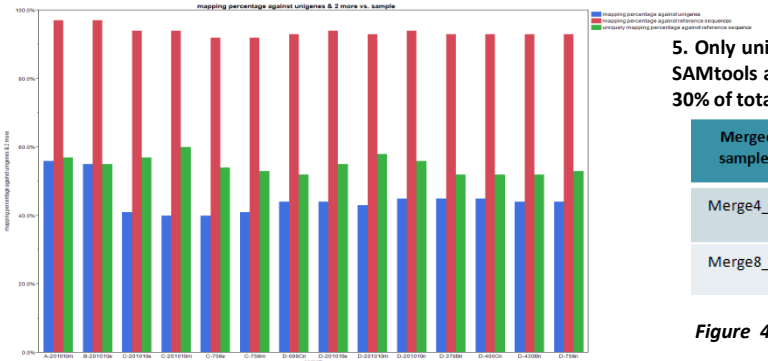

Figure 1. *Percentage of reads for each sample mapped against unigenes (blue bars), reference sequence (both multiple and unique hits - red bars; only unique hits – green bars). 20-1010, 7-56, 099C, 378B, 400C, 430B are the IDs of different trees; m – megagametophyte; e- embryo; n - needle.*

3. Capture breadth and depth were affected by higher multiplexing level (Fig. 2).

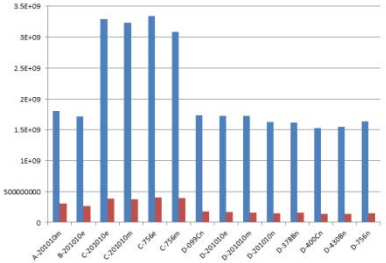

Figure 2. *Capture breadth (blue bars) and depth (indicated as sites equal to or above 5 fold coverage, red bars) against reference sequence.*

4. Reads that mapped to the reference sequence were merged for SNP detection and genotyping in the following pools: A and B (merge2), all the samples in C (merge4), all the samples in D (merge8) (Fig. 3).

Merged samples	Total reads (million)	Mapped reads (million)	Uniquely mapped reads (million)
Merge2	123	119 (97%)	69 (56%)
Merge4	404	376 (93%)	226 (56%)
Merge8	342	320 (93%)	184 (54%)

Figure 3. *Total and mapped reads numbers in the merged samples.*

5. Only uniquely mapped reads were selected for SNP detection by SAMtools and Freebayes (Fig. 4). (Minimum read depth=10, at least 30% of total uniquely mapped reads contain an alternate allele).

Merged samples	SNP numbers (SNP density) by SAMtools	SNP numbers (SNP density) by Freebayes
Merge4_10	2,370,705 (0.328 SNPs/kb)	3,112,436 (0.441 SNPs/kb)
Merge8_10	2,810,893 (0.396 SNPs/kb)	2,719,364 (0.422 SNPs/kb)

Figure 4. *Number of SNPs detected by SAMtools and Freebayes.*

6. Figure 5 demonstrates how numbers of genotyped SNPs decrease in pools of four and eight samples with a read depth of at least 8 reads per SNP per sample.

Merged samples	SNP numbers by SAMtools	SNP numbers by Freebayes	Intersection of two SNP caller
Merge4	169,769	352,973	154,488
Merge8	34,627	61,119	29,893

Figure 5. *Number of SNPs detected in merge4 and merge8 pools with read depth of at least 8 reads per each individual tree in the pools.*

Conclusions

1. Multiplexing strategies worked well for capturing targeted sequences and SNP discovery.
2. Higher multiplexing level reduces the coverage of each sample, but still provides high number of SNPs for efficient genotyping.
3. Sequencing depth for each sample can be increased by decreasing number of targeted genes.

Acknowledgements

We thank Dr. Chun Liang (Miami University, Oxford, OH) for providing the unigenes, Dr. David Neale , Dr. Jill Wegrzyn and Hans Vasquez-Gross (UC- Davis, Davis, CA) for providing the draft loblolly pine reference sequence and bioinformatics assistance , Dr. Matias Kirst and Leandro Neves (University of Florida, Gainesville, FL) for help with the SureSelect protocol, and the USDA NIFA for funding.