SNP DISCOVERY IN TROPICAL AND SUBTROPICAL PINES USING REDUCED REPRESENTATION SEQUENCING METHODS

Colin Jackson[1], Nanette Christie[2], Madison Caballero[3], Melissa Reynolds[2], Christopher Marais[2], Erik A. Visser[2], Sanushka Naidoo[2], Gary R. Hodge[1], Ross Whetten[1], Fikret Isik[1], Juan J. Acosta[1], Jill L. Wegrzyn[3], Alexander A. Myburg[2]

[1]Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, NC; [2]Department of Biochemistry, Genetics and Microbiology, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, South Africa; [3]Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT

This study performs SNP discovery and characterization using RNA-seq and targeted sequencing for use in developing a high throughput genotyping assay for tropical and subtropical pine species. Targeted sequencing was performed on six species of pine: *Pinus patula, Pinus tecunumanii, Pinus oocarpa, Pinus greggii*, *Pinus caribaea* and *Pinus maximinoi*. Sequence data was generated from a custom set of 40K capture probes (RAPiD Genomics Gainesville, FL) of which 30K were designed from single copy locations in v2.01 of the *Pinus taeda* genome assembly and 10K were designed from the P. *tecunumannii* and P. *patula* transcriptome assemblies. A total of 81 pooled samples were sequenced among the six species. The 81 pools represented between 4-8 trees from a single provenance and covered the natural ranges of the species in Mexico and Central America. Target sequencing generated between 3.1 and 7.7 million reads per pool with coverage of 20-30X across capture regions. Approximately1.1 million SNPs were detected in at least two of the 81 provenances, of which 403K are shared among most species. RNA-seq data was generated for the species mentioned above minus P. *caribaea*. Pooled RNA was isolated from shoot tissue of between 8-16 seedlings from two or more families per species. Paired end sequencing generated between 29.4 and 67.7 million raw reads per pool. Reads were trimmed and mapped to each species' respective transcriptome assemblies. SNP detection yielded between 426K and 1.3M SNPs per species. SNP probe design resulted in 1.8 million candidate probes designed between species and across platforms. The probes generated from each dataset were further assessed for unique vs. repetitive mapping against the v2.01 P. *taeda* genome assembly and similarity across species. Assessment of RNA-seq derived probes showed a large proportion of probes being unique to a given species with few being shared between. Approximately 53% of probes mapped to a unique location in the reference assembly. Target capture derived probes showed a larger proportion of probes being shared across species with greater than 80% of probes mapping to a unique location. From these 1.8 million probes, 323K RNA-seq and 121K target capture derived probes were selected for assessment on a screening array. Currently, 480 samples have been submitted for genotyping.