PINEMAP Year 2 Interim Report 2
November 2012
Aim 3—Genetics

Outcomes/Impacts

Describe how Aim-level activities, results, findings, techniques, or products contribute to project-level outcomes and impacts.

Aim 3 activities will make two contributions to project-level outcomes and impacts. One will be a genetic deployment tool to aid stakeholders in making sound genetic deployment decisions under changing climatic conditions, based on analysis of productivity data from test plantings across the region. The second will be to guide tree breeding programs and advance the understanding of genetic variation in resilience to climate variables in pine by using molecular markers to discover genes associated with important mitigation and adaptation traits, and characterize variation in pine breeding populations. Aim 3 will work closely with Aims 1 and 2 to provide input and benefit to tree breeding programs based on insights gleaned from the stand-level studies and modeling, to make tree improvement more efficient. Simultaneously, knowledge of traits varying at the gene level should contribute to the ecophysiology studies.

Outputs

List ***Products*** developed/completed January 2012-current (including published, in press, or in review peer-reviewed publications; other written materials such as white papers, research summaries, fact sheets, or popular press articles; audio or video products; etc.).

Aim 3 Peer reviewed publications (January-November 2012)

Aspinwall, M.J., S.E. McKeand, and J.S. King. 2012. Carbon sequestration from 40 years of planting genetically improved loblolly pine across the Southeast United States. *Forest Science* 58(5): 446-456. doi: http://dx.doi.org/10.5849/forsci.11-058

Holliday, J.A., T.L. Wang, S.N. Aitken. 2012. Predicting adaptive phenotypes from multilocus genotypes in Sitka spruce (*Picea sitchensis*). *G3: Genes, Genomes, Genetics* 9:1085-1093. doi: http://dx.doi.org/10.1534/g3.112.002733

Summarize ***Events/Activities*** (January 2012-current) as follows:

→ Provide a bulleted list of presentations (oral and poster) given at meetings or conferences. The format for citing presentations is as follows:
Presenter(s)/Author(s). Date. Name/title of meeting/conference, location.

Chhatre, V., T. Byram, D.B. Neale, J.L. Wegrzyn, and K.V. Krutovsky, January 14-18, 2012. Genome Wide Analysis of Genetic Associations with Environmental Variables in East Texas Loblolly Pine (Pinus taeda L.). Plant & Animal Genome XX . The International Conference on the Status of Plant and Animal Genome Research, San Diego, CA, USA

Krutovsky, K.V., January 14-18, 2012. Forest Genomics for Mitigating Climate Change and Breeding Resilient Trees. Plant & Animal Genome XX . The International Conference on the Status of Plant and Animal Genome Research, San Diego, CA, USA

Farjat, A., May 10, 2012. Climate variables and height growth in the Plantation Selection Seed Source Study field trials. NC State University Cooperative Tree Improvement Program Advisory Board Meeting, Raleigh, NC.

Whetten, R. May 10, 2012. Preliminary results from Genotyping-By-Sequencing experiments in loblolly pine. NC State University Cooperative Tree Improvement Program Advisory Board Meeting, Raleigh, NC.

Chhatre, V. May 22-23, 2012. Evolutionary and population genetics of the Lost Pines: Lessons from history. Western Gulf Forest Tree Improvement Contact Meeting, Bastrop, TX

Krutovsky, K. May 22-23, 2012. How genomics can help restore the Lost Pines. Western Gulf Forest Tree Improvement Contact Meeting, Bastrop, TX

Bawa R. and J. Holliday. July 6-12, 2012. Signatures of natural selection and local adaptation in *Populus trichocarpa* and *Populus deltoides* along latitudinal clines. Joint Congress on Evolutionary Biology. Ottawa, Ontario, Canada.

Holliday, J.A. and L. Zhou. July 7-11, 2012. Toward whole-genome association mapping and landscape genomics in trees. Botany 2012 – The Next Generation (Annual Meeting of the Botanical Society of America), Columbus, OH, USA.

Krutovsky, K.V., Oct 3 – 5, 2012. Forest Genomics for Mitigating Climate Change and Breeding Resilient Trees. Keynote presentation at the AdapCAR and IUFRO (WP 2.02.00) meeting Genetic aspects of adaptation and Mitigation: forest health, wood quality and biomass production, Riga, Latvia.

Krutovsky, K.V., Oct 16 – 17, 2012. Forest ecogenomics and genomic selection for mitigating climate change and breeding resilient trees. Final Conference and Workshops of Noveltree Project. Tree Breeding, Genomics and Evolutionary Biology: New synergies to tackle the impact of climate change in the 21 st century, Helsinki, Finland.

Whetten, R. Oct 24, 2012. Rust resistance screening and new molecular marker methods for pine. NC State University Cooperative Tree Improvement Program Contact Meeting, Asheville, NC

Isik, F. Oct 25, 2012. Universal Response Function development from the Plantation Selection Seed Source Study field trial data. NC State University Cooperative Tree Improvement Program Contact Meeting, Asheville, NC

Holliday JA, Zhou L. October 31, 2012. “Toward genome-wide association mapping in a model tree.” Texas Tech University Department of Biology seminar. Lubbock, TX.

Holliday JA, Zhou L**.** October 15, 2012. “Toward genome-wide association mapping in a model tree.” University of Maryland Department of Plant Science seminar. College Park, MD.

→ Provide a short narrative describing any workshops, courses, and/or trainings conducted.

Fikret Isik of the NC State University Cooperative Tree Improvement Program co-taught a week-long workshop entitled “Genetic Data Analysis: Applications in Plant and Animal Breeding” at the Czech University of Life Sciences in Prague, Czech Republic, June 11 – 15, 2012. The course covered statistical methods for analysis of field trials, including methods appropriate for trials planted at multiple sites to estimate impacts of climate variation on forest productivity. The course was jointly organized by NC State University Forestry Education and Outreach Program and forestry faculty from the Czech University of Life Sciences.

Konstantin Krutovsky of Texas A&M University chaired and co-taught a week-long workshop entitled “ Population genetic and genomic approaches to mitigate global climate change impacts on forest genetic resources and to breed more resilient trees” organized by Büsgen Institute, Department of Forest Genetics and Forest Tree Breeding, University of Göttingen and funded by EvolTree (Evolution of Trees as Drivers of Terrestrial Biodiversity, EU-funded Network of Excellence), September 24-28, 2012, Göttingen, Germany. The workshop taught how the application of population genetics and genomics can help to tackle global climate change challenge and guide forest protection and conservation programs aimed to mitigate its undesirable effects. It also demonstrated how modern genomic selection can be used for breeding more resilient trees that will be productive in harsh and variable environments. The workshop also covered population genetics and genomics tools and methods that can be used to better characterize available forest tree genetic resources. (http://www.evoltree.eu/index.php/component/jevents/icalrepeat.detail/2012/09/24/202/-/- ,http://www.uni-goettingen.de/en/evoltree-summer-school-in-goettingen/361359.html)

→ Provide a short narrative describing experiments or surveys conducted and/or analyzed.

Milestones and Work Plan Progress

Provide a short narrative describing progress and accomplishments on the year 2 milestones and work plan tasks listed below. Please also describe any changes to the Aim 3 milestones and/or work plan.

Year 2 Milestones

→ Version 1 of genetic deployment tool with provenance information (October 2012).

1) Texas A&M: Analysis of 73 tests of the WGFTIP Seed Source Study of 43 selected families (divided into two series) has been conducted with focus on using 1) minimum temperature of the coldest month (MMIN) and 2) summer heat/moisture index (SHM) as dependent variables and 1) height growth, 2) diameter and 3) volume as independent variables. MMIN corresponds to the most extreme event, and strongly correlates with the other minimum temperature variables investigated ($P < 0.0001$). The trend for MMIN was clearly linear and closely related to latitude. SHM showed a relationship with longitude that could be interpreted as curvilinear or perhaps indicating a threshold between two different distributions occurring at approximately 93.5° W. Highly significant response surface models have been developed; however, they pose further questions regarding the large portion of variation that is not explained by the weather explanatory variables alone. All raw progeny test data from this study has been supplied to University of Florida for the Tier 1 database.
2) NC State University: Growth (height at age 8) and climate data were analyzed for 19 field test sites of the Plantation Selection Seed Source Study using a multiple regression approach that tested a total of 68 climate variables and interaction terms for significance, to create a Uniform Response Function for the NCSU pine breeding populations. After refinement the model included 19 variables and interaction terms, and explained almost 68% of the variation in height growth in the dataset. Cross-validation to test the predictive accuracy of the model found predictive power was low, indicating that the problem is ill-conditioned due to co-linearity in the climate variables. Further refinement of the statistical model will use regularization or dimensional reduction methods to deal with the co-linearity in explanatory variables in an attempt to increase the predictive power of the model. Including additional independent variables related to site quality in the modeling effort is likely to increase the predictive accuracy of the model, provided that suitable data can be obtained for the test sites. The progeny test data from the 19 field sites used in this analysis have been provided to University of Florida for the Tier 1 database.

→ Determine appropriate genome reduction methods for genotyping by sequencing (October 2012).

1) Texas A&M: Agilent's SureSelect Target Enrichment System was used to select exon regions in loblolly pine. 647,634 baits were designed using 35,550 unigenes and the Agilent's eArray software. The total capture library size was approximately 78MB. Subsequently, genomic DNA of a haploid megagametophyte and a diploid embryo

from the same seed from the reference tree 20-1010 were isolated. After fragmenting the total genomic DNA into 150 to 200 bp long fragments, two sequencing libraries were built, one from each source of genomic DNA, and then hybridized with the capture library. After hybridization, the captured DNA was isolated using magnetic beads and amplified with index primers. The amplified selected sequences were sequenced using a Hiseq 2000 sequencer. After trimming and discarding low quality reads, sequences of 6.4 Gbp and 5.8 Gbp were obtained for the megagametophyte and embryo libraries respectively. These were *de novo* assembled into 184,907 contigs representing 58Mb. Blastn was used to check the *de novo* assembly against the draft 18Gb loblolly pine genome with the result that only 5 contigs failed to give hits. Attempts to map the sequence to the draft genome using CLC are currently underway. After filtering, this method has identified 51,623 SNPs or approximately one SNP for every 1,119 bp. Less than 8% of the initial unigene set is not represented. Foliage samples were collected from 397 individuals represented in the ADEPT II Association mapping population for the next part of the project.

2) NC State University: A published procedure for Genotyping-By-Sequencing using two restriction enzymes to fragment genomic DNA (Poland et al, PLoS One 7: e32253, 2012) was used to prepare DNA samples from two parents and 90 progeny of loblolly pine. Results were good - each of two Hiseq lanes yielded 175 million paired-end 100-nt sequencing reads, and analysis using the Java-based GBS tools in the TASSEL software package yielded from 47,131 to 103,669 candidate marker loci depending on software parameters. These candidate loci have been compared to the draft loblolly pine reference genome assembly produced by the PineRefSeq genome sequencing project to estimate the proportion of single-copy sequences. About 70% of the candidate loci mapped to the draft genome assembly, and of those, about half mapped to a single location, yielding over 16,000 candidate single-copy genetic markers. A linkage map with 12 linkage groups, containing 529 markers distributed over a total map distance of 1281 cM, was constructed to confirm the Mendelian segregation of a sample of these candidate markers. An additional cluster of markers is found in a small linkage group that is not yet connected to one of the 12; addition of more marker loci to the mapping dataset is expected to allow this fragment to converge with one of the 12 larger groups.
3) Virginia Tech: Genome complexity reduction by the 'RAD-seq' approach was tested for 10 haploid megagametophyte and two diploid needle samples. An initial test was completed in an attempt to optimize the stoichiometry between adapter sequences and available genomic DNA fragments from the restriction digest. Specifically, the P1 adapter was titrated between 0.01 and 1x relative to the protocol modified from Miller et al (2007) (https://www.wiki.ed.ac.uk/display/RADSequencing/Home). Once we were satisfied based on the bioanalyzer results of these titrations, we completed library preparation of the indexed samples were sequenced on a single lane of an Illumina MiSeq instrument (2x100 PE format). Results from this sequencing run were not encouraging – much of the data generated was from adapter sequences, indicating an excess of concatenated adapters with no genomic DNA inserts. Discussions with a variety of colleagues revealed that this is a common issue with

RAD-seq, and due to the success of the two-enzyme GBS approach being employed at NC State University, we decided to discontinue efforts to optimize RAD-seq and re-direct our efforts to the two-enzyme method described in (2) above.

→ Comparison of methods:

1. As noted above, the RAD-Seq method was dropped from further consideration. No data have been obtained for DArT-seq, but that procedure differs only minor respects from the two-enzyme GBS protocol tested, so the key comparison to be made is between SureSelect and double-digest GBS.
2. Cost per sample for the SureSelect method as implemented was about ten-fold more than for GBS
3. Recovery of polymorphic SNPs was comparable; SureSelect identified about 8% more total SNPs than GBS identified candidate markers (51,623 total vs 47,131 candidates)
4. Ability to distinguish paralogous variants from allelic variants has not yet been tested for SureSelect; 35% of GBS candidate marker sequences map to single locations on the draft genome assembly
5. Suitability for high-throughput sequencing is similar for both methods, although SureSelect is more technically complex
6. The degree of focus on transcribed regions is greater for SureSelect; virtually all candidate SNPs are in or very near transcribed regions or sequences similar to transcribed regions, while 17.4% (8199) of the GBS candidate markers align to sequences in the reference pine transcriptome assembly.

Year 2 Work Plan Tasks

→ Agree to standardized traits for analysis and a methodology for stratifying locations (April 2013).

Analyses to date for development of Uniform Response Functions have been carried out independently by the three tree breeding programs, to avoid infringement on the rights of cooperators to the proprietary datasets held by those breeding programs. Discussions are needed among the project investigators to decide if a joint analysis of all data together would be beneficial. If so, then consultation with cooperative directors and advisory board representatives will be essential to develop a mechanism for merging data across cooperatives for a region-wide analysis while maintaining the intellectual property rights of cooperative members.

→ Continue Phenotyping for: insect and disease resistance; response to fertilizer; insect and disease resistance (November 2012).

Measurements of growth in field tests are typically taken in the fall and winter, during the dormant season.

→ Reduction of Genome Complexity for generating molecular markers (April 2013).

1. Texas A&M:
    a. Attempts to map the sequence reads to the 18 Gb draft loblolly pine genome will continue. Protocols to maximize efficiencies will continue to be refined through 1) redesigning baits and 2) improved multiplexing of samples. DNA will be extracted from the 397 individuals in the ADEPT Association mapping population in preparation for genotyping. This is to support gene discovery through the analysis of the geographical distribution of alleles.
2. NC State University: Additional refinement of the double-digest GBS method will include testing an automated size-selection step to increase the reproducibility of recovery of genotypes at specific loci across multiple independent library preparations. DNA extraction and genotyping of tissue samples from progeny tests will be initiated. Customization of the open-source software used to analyze GBS datasets will be undertaken to improve the efficiency of data analysis and provide better integration with breeding program data management methods.
3. Virginia Tech: As noted above, the results from the RAD-seq approach were not encouraging. With colleagues at Texas A&M and NC State, we have therefore decided to focus on the two-enzyme approach and sequence capture. To improve the genomic coverage of data obtained from GBS, we plan to complete an additional test using non-methylation sensitive enzymes. We expect this method to provide data in non-repetitive regions, without a bias with regard to gene content. This will provide complementary data to that being generated by NC State using methylation sensitive enzymes (i.e., enriching for gene regions) for the purpose of genomic selection in loblolly pine pedigrees.

→ Continue Phenotyping for response to fertilizer (April 2013).

Broad Impacts & External Collaborations

Provide a short narrative describing broad impacts (i.e., far-reaching and possibly unanticipated outcomes resulting from Aim work, including contacts/collaborations with entities outside of PINEMAP).

The PineRefSeq CAP funded to obtain reference genome sequences for loblolly pine, sugar pine, and Douglas fir has been generous in releasing a draft loblolly pine genome sequence assembly for use in analysis of genotyping data generated by PINEMAP. The PineRefSeq project is also working to expand the loblolly pine transcriptome. Newly discovered genes will be used if we develop a new set of baits for exom capturing. Co-PI Loopstra is involved in the PineRefSeq transcriptome project and the exom capturing being done at TAMU. PINEMAP researchers will continue to collaborate with the PineRefSeq project team to coordinate storage and analysis of genotype and phenotype data obtained during the PINEMAP project, with the objective of helping to evaluate and improve the reference genome assembly produced by the PineRefSeq project and provide additional genotypic data sampled from a diverse set of loblolly pine genotypes. As part of this collaboration, Ross Whetten will make a presentation at the

PineRefSeq workshop at the Plant and Animal Genome Conference in Jan 2013 on the use of the pine reference genome sequence in applied tree breeding.

Contact is continuing between PINEMAP (represented by Ross Whetten, NC State University) and FoResTTraC, a European Union project to develop coordinated research plans to study adaptation of forest trees to climate change. The name is an acronym for "Forest ecosystem genomics Research: supporTing Transatlantic Cooperation", and the objective is to coordinate activities among North American and European researchers in the fields of ecology, genetics, genomics, and evolution of forest trees. More information about FoResTTraC is available at http://www.foresttrac.eu/. This collaboration has resulted in submission of a manuscript to Global Change Biology, entitled "Evolutionary response to climate change – evidence from tree populations", co-authored by scientists from Canada, USA, Finland, France, Spain, and Switzerland.

Training

Please list undergraduate and graduate students, postdocs, and technical personnel trained under this project and include a description of their research focus and/or role in the project.

- Rajesh Bawa, Ph.D. student, Virginia Tech. Rajesh is conducting Phenotyping of adaptive traits (i.e., cold hardiness and bud phenology) in the loblolly pine progeny trials and working with universities across region to conduct genome-wide association mapping between these traits and the re-sequencing data.
- Alfredo Farjat, Ph.D. student, North Carolina State University. Alfredo is carrying out joint analysis of progeny test data and climate data for the Cooperative Tree Improvement Program.
- Tomasz Koralewski, Postdoctoral Research Associate, Texas A&M University. Dr. Koralewski is building a location database for the progeny tests and the parents represented in those progeny tests that can be linked to climatic data, and performing URF analysis to optimize seedling deployment.
- Mengmeng Lu, Ph.D. student, Texas A&M University. Mengmeng is evaluating efficiency of the Agilent SureSelect target gene enrichment system for Illumina GAII and HiSeq high-throughput sequencing instruments, based on capture probes designed using the Agilent eArray tool and the most recent and comprehensive unigene set consisting of 35,550 sequences representing individual genes expressed in loblolly pine.
- Will Kohlway and Ben Rusche, undergraduate student research interns, North Carolina State University. Will and Ben are working on pilot scale experiments to test genotyping-by-sequencing as a cost-effective means of genotyping the experimental populations.
- Laura Townsend, M.S. student, North Carolina State University. Laura's research focus is genetic variation in loblolly pine for adaptability to climate variables.
- Jianxing Zhang, Ph.D. student, University of Florida. Jianxing is conducting analyses linking progeny test performance with climatic data and developing a database for the URF analysis to optimize seedling deployment.

Collaborations and "integrated" knowledge developed

Provide a short narrative describing new ideas, research questions, or insights that have arisen through work and discussions with colleagues, stakeholders, and others. In addition, explain the extent to which you intend to incorporate this into PINEMAP milestones and/or your Aim work plan.

Needs from/linkages to other Aim groups

→ Provide a bulleted list outlining research results, data, products, or assistance that your Aim group needs from another Aim group.

- Aim 1
    - Coordination of sampling strategy, so that tissue can be collected from all individuals on which individual-tree measurements of any sort are made during the course of this project, for eventual genotyping. The storage solution developed at NC State allows ambient-temperature storage of foliage or of cambium disks taken from increment-core wood samples for periods of days to weeks, reducing the burden on field crews taking other measurements or samples.
- Aim 6
    - Information on climate science and future climate scenarios for decision support system module development
    - Presentation of state of climate science to cooperative members
- Aim 4
    - Information on regional changes in insect risks
    - Impact of policy changes on reforestation needs
- Aim 5
    - Course for graduate students to take on conducting trans-disciplinary research

→ Provide a bulleted list outlining research results, data, or products that your Aim group has compiled that have value or relevance to another Aim group (and note which Aim group).

- Aim 6
    - AIM 3 inputs and outputs provided to AIM 6 group (Ryan Boyles and Heather Dinon) as part of the aim interactions analysis for the purpose of Decision Support System (DSS)

→ List any additional potential linkages to other Aim groups.