Conifer Translational Genomics Network CAP
2010 Annual Meeting—June 17-18, 2010

Report by Advisors Luca Comai, Julie Ho, and James Johnson
Responses (inset text) from the CTGN CAP Team

The CTGN group is distinguished by its commitment to crop improvement and by its breadth of perspective from genetic discovery to product development and deployment. The group continues to progress in genetic and phenotypic data collection of loblolly pine and Douglas-fir. A sufficient body of information is available now to start developing a comprehensive summary of their results in the project's final year, particularly with regard to validation of marker-trait associations. The motivation for this would be three-fold: (i) to fulfill a primary objective of the original proposal, (ii) to demonstrate more systematically the potential value of genetic markers in applied breeding programs, and (iii) to broaden inference on the relative efficiency of different approaches to marker-based breeding and molecular genetic data analysis in general. While implementation of marker-informed breeding (MIB) may fall beyond the scope of this project, the Advisory Board recommends preliminary discussion of logistics such as assay technology, marker conversion and funding for individual breeding programs to genotype future operational tree improvement populations at validated QTL when appropriate.

1. Validate SNP by quantitative trait associations (discovered under prior USDA and NSF funding) in operational tree improvement populations of loblolly pine, slash pine, and Douglas-fir.

The CTGN team has developed an excellent set of ~5000 SNP suitable for high throughput genotyping in loblolly pine. A comparable set of SNP has the potential to emerge in the next year for Douglas fir. The coordination between the different groups in development and genotyping has been very good and as a result these tools and data sets can be applied to the validation effort. Success rates average >80% and 68% for DNA sampling and genotyping, respectively. The process of associating SNP to trait has met early success in that several candidate SNP are emerging for traits such as tree volume and height. These are traits that are more challenging but also more important than the high heritability traits initially considered in association studies. At the same time the effort has met challenges because it is not clear yet that high value SNP are available for broad scale adoption.

Recommendation: The committee recognizes that SNP discovery in Douglas-fir was not an objective of the grant. Yet, the technological scenario has changed since the application was written and given to advances in next generation sequencing SNP discovery is now considerably cheaper. In this context, and also considering the uncertainties on SNP number needed for efficient association, the number of SNP developed for Douglas-fir should be higher than the 1,500 currently planned. The choice of genotypes and tissues, the rationale for the diversity set and the strategy for defining high quality and informative SNP should be considered very carefully.

For Douglas-fir SNP discovery, we chose the genotypes and tissues based on the following considerations. For the JGI 454 sample, we chose to use a single

genotype from one of the included breeding populations so that we would have some information that would allow us to distinguish alternative alleles from alternative loci (i.e., there cannot be more than two sequences from a single locus in this sample). Because of the timing of the JGI project, RNA was collected from multiple tissues on a single day. The limitation to this approach is that we may miss genes that are expressed under other conditions (e.g, times of the year), and will miss much of the genetic diversity in coastal Douglas-fir. Therefore, we used 454 and Illumina sequencing to sequence a second pooled sample consisting of equal amounts of RNA collected from many genotypes and tissues throughout the year (diversity sample). Although we collected some tissues from mature trees that were differentiating flowers, we concentrated on seedlings so that we could have all trees growing in a common environment and could impose cold and drought treatments. To obtain high genetic diversity, we used seedlings from 79 seed sources. These seedlings, which were derived from breeding program seed orchards, were collected at a seedling nursery where these trees were being produced for outplanting. We feel that the diversity sample is particularly appropriate for meeting our objectives for the following reasons. First, because we are particularly interested in genes that confer resistance to cold and drought, we felt it was important to increase the representation of these genes by imposing cold and drought treatments. Second, because our objective is to develop SNPs and SNP chips for the coastal Douglas-fir tree breeding community, the 79 seed sources we chose were the best sources of material to represent Douglas-fir breeding programs in Oregon and Washington. Ultimately, our inference population is eight breeding populations that were derived from about 33,000 parent trees collected from the wild. We were asked whether the SNPs we develop will be useful in this target population. We think the answer is yes because (1) the diversity population sampled the breeding populations of interest, (2) gene flow is extensive in Douglas-fir, leading to relatively little population-specific genetic diversity (at least for the genes studied so far), and (3) the approach of developing Douglas-fir SNPs in one set of genotypes, and then applying them in others has already worked well, e.g., Eckert et al. (2009).

Our approach for identifying high quality SNPs has not been completely worked out, but will rely heavily on having extensive sequence depth. We will combine ~2.9 million 454 reads with ~92 million Illumina reads (diversity sample and other samples generated in collaboration with R. Cronn) to balance the advantages of each approach (e.g., types of sequencing errors) and to

ensure we have extensive coverage (depth) for the high-quality SNPs we call. Ultimately, high-quality SNPs will be called based on sequence depth, read mapping quality, nucleotide call quality, and the presence of SNPs in alternative sequencing platforms.

Foliage tissue is readily available and can be collected without concern for damaging valuable trees in seed orchards or progeny tests, unlike xylem or root tissue samples. Normalization will help to reduce the great disparity in mRNA abundance typically found in photosynthetically active tissues between the mRNAs that encode proteins involved in photosynthesis and other cellular mRNAs, and should allow SNP discovery to sample an array of at least 10,000 different transcripts with sufficient depth of sequencing on an Illumina instrument. The diversity set to be sampled for SNP discovery purposes will include ten high-breeding-value parents from each of ten breeding zones, to sample SNP diversity both within and among breeding zones. Methods for identifying SNPs in sequences from pooled samples have been reported recently (Bioinformatics 26:i318-24, 2010), as have improved methods for identifying high-quality SNPs (Bioinformatics 26:1029-35, 2010). We will explore the available tools and determine how best to structure the SNP discovery sequencing experiments to fully exploit the capabilities of available software tools.

Recommendation: A model that incorporates map information and realistic LD values for intra and interfamilial comparisons should be useful in ascertaining how many SNP are necessary and sufficient for efficient discovery.

Simulated datasets analyzed using GS3 show that a training population with reliable phenotypes and genotypes allows accurate predictions of genetic value for related individuals, utilizing the LD from recent generations of crosses, provided that the prediction population has sufficient shared ancestry with the training population. These simulations will be expanded to include pedigree structures similar to those found in the Douglas fir breeding program to determine the number of SNPs that would be suitable for prediction of breeding value for various levels of kinship. Very little is known of LD at a chromosome level (i.e., for distances greater than 3 kb) in any conifer, so assessment of what constitutes "realistic LD values" will be based on the assumption that published observations for LD over kilobase distances can be extrapolated to the level of whole chromosomes.

Recommendation: As time permits, evaluate alternative methods of allele calling (identity-in-state, haplotype, identity-by-descent). It may be helpful to compare detection power and analysis

results using a bi-allelic (IIS) versus multi-allelic (haplotype, IBD) model.

Map locations for the SNPs genotyped in the loblolly pine samples must be determined in order to explore haplotype-based approaches, but we have begun using IBD approaches and preliminary results are encouraging for training and prediction populations with substantial coancestry. Efforts are underway to exploit the pedigree information for genotyped individuals to improve the quality of SNP calling in the Illumina GenomeStudio software, and mapping analyses will be used as a quality-control check on the resulting genotype calls. Accurately called SNP genotypes should provide good-quality genetic linkage information and allow construction of a map that includes many of the genotyped SNPs. Determining the true order of tightly-linked SNPs will require the use of software designed for human genetics, because we don't have a single full-sib family large enough to allow us to reliably order markers at sub-centimorgan ranges using software designed for inbred crop analysis. These analyses will be very time-consuming, because only a handful of markers can be analyzed at a time, but will allow much more accurate imputation of missing data within populations that have significant haplotype structure. Populations of unrelated trees are not expected to show any haplotype structure based on reported observations of individual genes, but this hypothesis has not been extensively tested in large samples with many markers.

Recommendation: Develop and perform standard validation procedure across all available estimation sets (CTGN, ADEPT/2, ADAPT, meta-analysis thereof): (i) compute accuracy of prediction or selection for (a) predicted vs observed (retrospective study) and for (b) predicted vs 'truth' (simulated on realized population parameters), (ii) compare power of different statistical models, and (iii) compare utility of different methodologies (association mapping, whole-genome evaluation) for different trait complexities.

Cross-validation approaches will be developed and applied to the available datasets to determine prediction accuracy based on observed data, per item (i,a). Comparison of predicted values to values simulated based on realized population parameters (item i,b) seems more likely to provide information about the accuracy of the assumptions made during the simulation process than information about the accuracy of the analytical methods used. We have already observed, when using Bayesian methods for prediction, that results from analyses are more accurate when the priors using for Bayesian analysis happen to match the distributional assumptions made during simulation. We take this observation to mean that no matter how accurately a particular analytical routine predicts value in simulated populations, the most meaningful

test of its value is its prediction accuracy using data from real populations with real (and therefore stochastic) distributions of parameters. Comparison of the power of different methods (item ii) is therefore best carried out on real data, although simulated data can be used to test a wider range of population parameters than are available in real data. The utility of different approaches (item iii) is heavily dependent on the kinship structure of the available training and prediction populations. This will be demonstrated using simulated populations with a range of kinship structures, and the simulation results validated using real data for a subset of the simulated structures. The simulations can also explore the dependence of accuracy on heritability for a wider range of heritability values than are available in real data.

Developing a standard method for cross-validation and making open-source tools for cross-validation available on the website is a good suggestion, and will be implemented. Metrics to evaluate the utility of predictions for applied breeding programs will also be developed, including rank change approaches as well as simple correlations.

The CTGN group originally proposed to validate marker-trait associations identified in experimental populations under the ADEPT/2 project. While this approach might have given the group a more substantial dataset for proof-of-concept, trait targets were changed later to meet industry demand. Because opportunities to validate associations from ADEPT/2 in CTGN datasets are therefore very limited, more emphasis on validation within data sets is required. Most, if not all, of the key analyses (ASReml, SAS/Mixed, TASSEL, GS3, R) for phenotypic data analysis, association testing, whole-genome evaluation, cross-validation or simulation already have been applied by at least one of the four breeding programs. Now, the group only needs to standardize and extend these analyses to all available data sets for broader inference.

Relative efficiency (or detection power) of different statistical models can be evaluated, including use of phenotypic BLUP, A or G matrix and polygenic or family effect. In addition to correlation of predicted to observed (or simulated) MEGV (or SNP BV), similarity coefficients (e.g. top 25% or bottom 25% ranked individuals) could be computed to indicate accuracy of selection. The latter may prove more relevant to breeding programs than correlation alone.

New target traits, such as growth and form, are lower in heritability than many of the traits originally studied under ADEPT/2. While the advantage of MAS for low heritability traits surpasses that for high heritability traits in general, the method of association mapping, in particular, is not necessarily well-tailored to low heritability traits. A validation procedure that compares accuracy and resolution of association mapping to that of whole-genome evaluation, for example, might be informative in examining this potential issue.

Finally, it would be helpful to reviewers if the group would provide a genotyping timeline (and contingency plan if necessary) to assess feasibility of reaching target numbers (10,000 trees x 7,600 or 4,800 SNP) in the project's final year.

Genotyping for the UF project has been completed on the loblolly chip (BC1 population). All DNA sampling will be completed on the slash pine population (circa 3600 samples) in Sept 2010 and will be run on a 7600 chip at UC Davis. All UF loblolly pine genotyping should be completed by January 2011. SNP genotyping of all Douglas-fir trees will be conducted in early 2011.

2. Identify and economically evaluate methods for incorporating marker-assisted selection into conifer tree breeding programs.

Proof of economic feasibility will be critical for widespread adoption of MIB. The group has made a great start developing a powerful analysis tool which can demonstrate value of MIB to industry, that may be augmented by net present value considerations, and that can be used to optimize future methodology and design.

Recommendation: Consider combining Tree Genome Simulator + Simetar functionality for more robust and comprehensive analysis: (i) encompass multiple entry points (x SNP requirement), (ii) update metrics along with training set**,** and (iii) incorporate additional breeding program parameters, such as multiple trait targets, selection intensity, testing pattern and germplasm sharing.

The economic simulation is an important part of both objective 2 and objective 5. The plans are to clean up the user interface for the simulator to make it more robust when released as a publicly accessible tool. It is anticipated that this will include a concise list of inputs, with multiple case studies as examples, and a rudimentary user's manual. One of the strengths of the software is that it can accept input from any number of sources including the researcher's own particular biases. We should be able to demonstrate this point by including several case studies and examples with some of these based on output from Tree Genome Simulator. The link between programs can be made more explicit and easier to implement by either matching the formats between Output and Input tables, or providing a Perl script or SAS macro to convert the output of the genetic simulator into the required format for input into the economic simulator. The Simetar® Tree Breeding Evaluation Simulation currently includes multiple entry points and multiple econometric evaluation techniques that can be illustrated with the case study approach. It also includes the ability to build in assumptions about future cost efficiencies of technology, but does not include the ability to incorporate increased accuracy in selection as we add to future databases as suggested by the reviewers. Making this addition will be considered. Modifying the Tree Genome Simulator to include multiple trait targets and the other suggested features will require substantial time, and the limiting factor will be how much

of Jianbin Yu's time is available to work on programming as distinct from his work on the bioinformatics aspects of SNP discovery in Douglas fir.

Marker data generated for one purpose (e.g. individual selection) can be leveraged at several other stages in the breeding process, such as parentage analysis (quality control), use of marker-based relationship matrices in BLUP computation, family prediction/selection, and product placement (GxE). In practice, each subsequent year of genetic and phenotypic data will be added to the historical reference or training set. We expect that, as the number of records and the breadth of germplasm sampled increase, so will accuracy of prediction. These additional opportunities will enhance the potential value of MIB and could be incorporated into a more accurate economic risk analysis.

Simulation tools developed by CTGN may be used to improve our understanding of experimental design (e.g. within/between family selection, seedling/clonal testing, administrative units, marker selection). For a given trait complexity and starting sample (number of individuals, QTL frequencies), which population parameters (number of entries, reps, effective SNP, realized heritability, kinship or genetic similarity to target population, etc) can be optimized to make a dataset most efficient for prediction/selection in elite breeding populations? These findings no doubt would add value to crop improvement research beyond the conifer community. In addition, economic risk analyses conducted by CTGN can serve as a good example to other research groups that are not as well integrated with industry or the breeding and extension communities and therefore may lack this critical perspective on deployment.

3. Develop databases (TreeGenes) and web-based tools (Dendrome) to facilitate all aspects of the Project.

The group has done an excellent job at implementing the The Dendrome/TreeGenes and DiversiTree sites (http://dendrome.ucdavis.edu) represent an important resource for the forestry breeding community with a recorded average of ~7,000 hits per month. Funding for these resources after the CTGN grant ends is a concern but an obvious source would be the forthcoming loblolly genome project, especially if we expect that all trees sampled by CTGN will be sequenced in the future.

Recommendation: As time permits, enhance marker data delivery: (i) generate summary statistics by marker locus for user-defined sample (website), (ii) impute missing marker data, and (iii) share tools for data formatting. Summary statistics by marker locus would complement the quality control filters already in place for phenotypic data and may include MAF, %heterozygous, %inconsistent, X^2, PIC, allele number, etc. Imputation of missing marker data, based on haplotype analysis, may improve balance and detection power. Already developed at NCSU, SAS code might be linked to the CTGN website so that users can export marker data formatted for TASSEL, GS3, ASREML, SAS, etc. The latter would facilitate a common set of analyses by different groups on different data sets. Will modifications made to TASSEL, for populations with large numbers of SNPs, be made accessible to researchers beyond the conifer community through this website?

We agree that all suggested enhancements to the TreeGenes and Dendrome

databases would have significant value to the community. PD Neale continues to identify long-term funding sources for the databases as he has over the last 20 years. All software developed by the Dendrome database team is open source.

4. Develop an international conifer genetic stock center.

Molecular biology resources and regents for conifers developed over the years (cDNA libraries and clones, BAC libraries and clones, mapping and association population DNAs, and PCR primer sets) will be curated along with clonal tree archives to form a conifer genetic stock center to benefit breeders and geneticists. Continued funding of this resource may be alleviated if the industry is motivated to move into MIB.

It is unlikely that the private sector will support a genetic stock center. PD Neale seeks to identify a single source of funding for both the Dendrome database and the Forest Tree Genetic Stock center (http://dendrome.ucdavis.edu/FTGSC). The US Forest Service is the federal agency that should provide this support and we ask NIFA to call upon the Forest Service to do so.

5. Create and implement an education plan for undergraduate and graduate curriculum development in genomics-based breeding in forest trees.

The CTGN Team is to be congratulated for conducting the first short course, entitled Genomics and Tree Breeding in Forest Ecosystems, in Davis, CA in June, 2009. By all accounts the course was well-attended and the evaluation report indicates it was very well-received. Unfortunately the follow-up course for 2010 was canceled due to low enrollment (5 students from the U.S. and 13 overall). The Advisory Board recommends that the course be re-offered in the summer of 2011, and be updated in accordance with the suggestions from the 2009 participants. This is a valuable outreach activity for the CTGN group, and perhaps consider providing some scholarships for participants, particularly graduate students.

Agreed. We intend to offer the course in 2011, though the dates remain undetermined at this time. It will likely have to be later in the summer to accommodate the international symposium we will offer in June. We essentially provide the course free of charge to all attendees. We asked for a fee of $250 in 2009 to partially cover the cost of room and board (subsidized by the project). This modest fee is designed to insure that applicants are relatively serious about attending. We will consider offering a full scholarship that will cover the cost of travel and room/board for 2 graduate students based on need and application essay.

The CTGN Team appears to be taking the Education objective very seriously. We recommend that the Team continue to develop the online education modules that may be made available on

the web as "legacy" pieces from the CTGN Project. These modules can and should be used by instructors both here in the U.S. and abroad, in both undergraduate and graduate courses, and for continuing education for professionals as well.

Agreed. It is a major milestone for the next 2 quarters.

The Advisory Board recommends the CTGN Team create a table that displays all the graduate students who received full or partial support from the CTGN Project, and show pertinent information such as the student's name, institution, degree received or pending, thesis or dissertation title, publications, and current status. If working, state where. This should be retroactive back to the beginning of the project and should be updated as year four draws to a close. This would be helpful to show the full impact of the CTGN Project on the development of graduate students who will become the genetics workforce of the future.

We currently feature the names of such students at our website (http://dendrome.ucdavis.edu/ctgn/people) but without the detailed information that you suggest. We will develop a new table with the recommended information.

6. Develop and deploy an extension curriculum for continuing education in genomics-based breeding for practicing tree breeders and forest tree gene-resource managers.

The Advisory Board commends the CTGN Team on the re-tooling of the website and the creation of the newsletter. Both efforts greatly enhance the communications and outreach capabilities of the CTGN project, and show great responsiveness to recommendations from 2009.

The Advisory Board suggests beefing up the econometrics analysis for MIB, and perhaps consider teaming up with an economist to implement some standard tools for econometric analysis, such as Net Present Value, Internal Rate of Return, Composite Rate of Return, Annual Equivalent Value, etc. These standard tools for evaluating investments are well-understood by forest industry, and it appears that the economic benefits of MIB will be critical to moving the technology forward to application. If a beefed up economic analysis can be completed in time, we suggest including it as a key component of the symposium scheduled for June, 2011.

Agreed. The Simetar program has the functionality to deliver on all of these suggestions (see response at Objective 2). We are conversant with most of them ourselves, as they are common in the forest industry for evaluating tree improvement research and operations. We have put a place in our symposium program for delivery of this topic (the 'Hurdles to application' theme).

The Extension component that deals with direct interaction with Tree Improvement Cooperative members was not discussed much today. However, this interaction is a key part of the Extension function. The Advisory Board suggests creating a table that shows the individual meetings with Co-op membership for each of the participating Co-ops, the location, date, number of participants, and nature of the interaction, for example, titles of presentations, publications provided, etc. This will display the full interaction of the CTGN Team with the primary target

audience.

We believe the table we present at our website on all outreach presentations covers this recommendation (http://dendrome.ucdavis.edu/ctgn/educationextension/presentations.php). It has been updated since the June 2010 CTGN Annual Meeting.

References cited in CTGN responses

Eckert, A.J., Bower, A.D., Wegrzyn, J.L., Pande, B., Jermstad, K.D., Krutovsky, K.V., St. Clair, J.B., and Neale, D.B. 2009. Association genetics of coastal Douglas fir (*Pseudotsuga menziesii* var. *menziesii*, Pinaceae). I. Cold-hardiness related traits. *Genetics* **182(4)**:1289-1302.

Bansal, V. 2010. A statistical method for the detection of variants from next-generation resequencing of DNA pools. *Bioinformatics* **26**:i318-324.

Malhis, N. and Jones, S.J.M. 2010. High quality SNP calling using Illumina data at shallow coverage. *Bioinformatics* **26**:1029-35