Response to SAB Report

The Scientific Advisory Board submitted a thorough and timely report to our team following our annual meeting in Columbia, Missouri September 9-13. The report summarized discussions between our team and the SAB at the meeting.

We have carefully considered, and discussed at length, the recommendations made by the SAB team. We found all recommendations to have merit and believe they would add value to our projects outcomes. However, given our approved grant sideboards and guidelines, and our limited remaining budgets, we may simply not be in a position to implement all suggestions within the current grant.

In this document, we will restate the SABs recommendations (displayed topically, highlighted in red) and provide written responses to them (in blue font). These will be shared with all team members, all SAB members (including those who did not attend this year's meeting), and our NSF project managers. All documents will be posted at our project website, as in year's past (http://www.hardwoodgenomics.org/) .

III. BAC Libraries and Physical Maps

The project has put substantial effort into developing BAC libraries and physical maps for Northern Red Oak and Black Walnut. The libraries that have been produced appear to be of high quality, covering each genome at approximately 10X clone depth, with low organellar content, and producing clear results in hybridizations using single gene probes. However, efforts to produce a "gene space" physical map have not been promising because hybridizations with cDNA probes representing the transcriptome of Northern Red Oak have not revealed a substantial fraction of gene-poor BACs. Subsequent *in silico* analyses have demonstrated that the assembled genomes of peach, grape and *Populus* also contain few gene-poor regions, so that a strategy targeting gene-rich BACs would be unlikely to greatly reduce the complexity and extent of a physical mapping project. Despite the relatively negative results, the SAB believes that the data generated from these experiments is valuable, and merits a short publication on its own.

We agree that publication of results should be a priority here. We have debated for some time what a good paper would look like and where it should be submitted. Simply publishing a BAC library or two has little traction today, and we believe it needs to be enhanced to meet publishing standards. We are currently inclined to submit a single publication that includes the BAC libraries and the results of the experiments that we will conduct in accordance with the recommendations that follow. We believe this product will have broader appeal.

The budget will only allow for fingerprinting of approximately 30,000 BACs, which would yield a whole genome physical map of only 5X coverage. Based on past experience and a general consensus in the literature, this map would be highly fragmentary and therefore of limited utility. The investigators are therefore proposing an alternative use for these BAC libraries. They would target several genomic regions (approximately 1 Mb in total) using overgo probes designed from existing plant genomes. The proposed project would focus on regions identified in peach and chestnut that have been previously well-characterized. The SAB agrees that a targeted, hypothesis-driven BAC fingerprinting and sequencing experiment would be the best use of the remaining resources in this portion of the project. We have several recommendations in this regard:

1. The experiment should be driven by hypotheses grounded in the molecular evolution literature, and should address a key question that would be of broad interest (and therefore likely to yield a high impact publication).

2. Similarly, the experiment might also explore and/or demonstrate a new technique in physical mapping, genome assembly and/or haplotype reconstruction. For example, the idea of doing some deep PacBio sequencing on a BAC pool may have some merit, particularly if this has not yet been demonstrated. In this case, it might make sense to target some regions that have been recalcitrant to assembly in chestnut and European Oak.

3. The experiment should ideally be integrated into the other objectives of this project as much as possible. For example, at least a portion of the sequencing could be targeted toward BACs containing genes that show extraordinary responsiveness in the stress experiments.

4. If it makes sense in the context of the hypotheses generated above, the investigators should consider using other BAC resources at their disposal, such as the Yellow Poplar library that was produced at CUGI as part of a previous project. This would enhance the phylogenetic coverage of the experiment and perhaps lead to deeper insights.

We find these guiding recommendations, with perhaps the exception of point number 4, to be very consistent with ideas generated by our team. We have developed a brief study plan that will outline our experimental approach, timeline and budget that will permit completion in the final year of the project. This will be included at the end of this report.

IV. Marker Development

This project has done excellent work to develop broadly useful sequencing and marker resources for a number of understudied yet ecologically and economically important forest tree species. For example, the project has completed 45 Gb to 100 Gb of sequencing of transcriptomes of Northern Red Oak, Black Walnut, and Green Ash seedlings, thereby greatly enhancing coverage of the transcriptomes of these strategically-important species. Furthermore, shallow whole genome sequencing has been completed for most of the species in

Table 1, resulting in initial genome assemblies and fairly complete chloroplast genomes. Furthermore, SSR markers have been developed for most of these species by querying these sequence databases for SSR motifs, designing primers, and testing against a panel of parents and offspring for mapping populations. This work will result in several manuscripts describing the marker resources, and will provide valuable tools to future researchers investigating the genome structure and population genetics of these species.

Based on previous suggestions from the SAB, the project is now aggressively pursuing marker development using a "Genotyping by Sequencing" (GBS) approach. Initial results with a pilot study in Northern Red Oak are quite promising, yielding several hundred markers despite very low per-sample sequencing depth. The project has now established a collaboration with the BGI-UC Davis sequencing center to perform GBS for 250 Northern Red Oak full sibs from the main mapping/QTL population produced by the project. This approach appears to have a high likelihood of yielding thousands of mappable markers, thus fulfilling the stated project objective of producing a dense genetic map for this species. If successful, a similar approach will be taken for Black Walnut.

The remaining challenges appear to be on the bioinformatics side, since different pipelines have yielded quite different results with the initial data. The SAB recommends aggressive screening of these GBS loci based on depth, genotype quality, and Mendelian segregation, with the goal of establishing a framework map with reliable markers. At the same time, it is important to keep in mind that if markers are screened too aggressively for Mendelian segregation, it may cause bias against markers that display segregation distortion for biological reasons (e.g., prezygotic or postzygotic selection in the family), thereby causing gaps in the maps. It may also be useful to perform both *de novo* and reference-based genotyping, with the latter being based on assembly to the assembled Northern Red Oak transcriptome.

We strongly agree with SAB that GBS loci require aggressive screening. Our teams at Notre Dame and Clemson have invested significantly in putting in place a rigorous pipeline for data analysis and interpretation of rad tag sequences. We feel the experience of our staff in genetic mapping and bioinformatics gives us a real advantage on this score.

Based on our early results with GBS in red oak, we are moving ahead aggressively with developing a similar approach with black walnut (before the end of year 3), and we are giving strong consideration to use of this approach for out genetic maps of honey locust and green ash, in year 4. We are very thankful to the SAB for last year's nudge in this direction.

V. Population Development

This objective is proceeding well for the most part, with full-sib families fully assembled for Northern Red Oak and Black Walnut, yellow poplar, and green ash (through a collaboration with the US Forest Service). The development of a full sib mapping population for honey locust

has lagged, in part because of the large pool of pollen parents discovered in the open-pollinated seedling families during PE analysis. The SAB recommends that one more attempt be made at developing a large full sib mapping family based on seed collections from Fall 2013, after which it would be prudent to proceed with a half-sib family. We recommend that the same family be used for mapping as is being used for stress treatments this year, if possible. This family, derived from Ames number 10, has already produced a large number of seedlings that can be planted in Missouri or Tennessee following stress testing. Although the number of full sibs detected is small, this would not be a disadvantage if the plan is to proceed with half-sibs, since a large mix of pollen parents would be desirable for a half-sib family (i.e., to avoid biases due to unequal representation of pollen parents).

Agree, in part. We are indeed making one more stab at developing a full-sib family, but it is not going to be Ames #10. To recap, we started the project with a large seed collection from a single tree in Butternut Valley, TN. About 400 OP progeny were established in middle TN on a field site in anticipation of acquiring markers for paternity analysis. This parent tree failed to produce seed in years 2 and 3 of this project, and our fear was that we would not have enough trees to meet our FS needs. So, in year 4 we collected from 6 additional trees at Ames, TN. Markers have subsequently demonstrated we have relatively small FS families for all parents. Based on samples genotyped to date, it appears we may have as many as 60 or more FS seedlings from our original parent, in those trees residing in the field planting/freezer. Our collaborator in TN is collecting seed from this tree as this is being written. We are hopeful we can develop a mapping / QTL population that exceeds 100 trees this coming year, using the seed being collected this Fall, and complete our map.

Clonal propagation of full-sib Northern Red Oak populations has been very successful, and the SAB recommends that these be planted in replicated plantations in Missouri and Tennessee to allow assessment of Genotype x Environment interactions. Furthermore, clonal replicates will enable calculation of broad-sense heritability, which will help focus future phenotyping and mapping efforts on the most promising traits. Clonal propagation of the other mapping populations would also be desirable if it will fit within the existing budget and project timeframe.

Completely agree, and plans are underway to do just that. Arrangements are also being made for clonal propagation of the green ash mapping population, to be planted at two sites.

VI. EST and Transcriptome Sequencing

The project has been very successful in producing, assembling, and analyzing transcriptome data from a large number of species, tissues, and conditions. This has been a gargantuan task, and the resulting resources will be invaluable for this and future projects. Unfortunately, as is often the case with such projects, there have some difficulties in producing libraries for some of

the experiments. Most notably, the initial RNAs prepared from the cold, heat, drought, and wounding of Northern Red Oak and Black Walnut did not pass QC for library construction. Because of the central importance of these species to the goals of this project, and the potential importance to the larger community of researchers, the SAB recommends that these experiments be repeated.

The SAB also has the following recommendations for the stress treatment RNAseq experiments:

1. It is essential to barcode individual biological reps from these experiments so that these may be properly analyzed statistically to identify up- and down-regulated genes. At the very least this will enable the researchers to address the impact of pooling replicates and identifying up-regulated genes based on fold-change alone. Furthermore, this will likely enable publication in a higher impact journal due to the enhanced statistical rigor.
2. We strongly recommend that the researchers consider increasing the number of time points and/or treatment levels so that the physiological and transcriptional responses to these stresses can be more thoroughly characterized. Although it may be necessary to focus on fewer stressors with possibly less replication in order to accomplish this, we believe that this will ultimately result in more readily-interpreted results and therefore a stronger publication compared to the currently-planned experiments. The ozone experiments are a good model in this regard.
3. It is important to segregate control plants from stressed plants, at least for a subset of the experiments, due to the possibility of volatile signals from the stressed plants that will affect the transcriptomes of the controls. We recommend a minimum of two meters between treatments.

We find the SAB's recommendations to be compelling and desirable. The issue of library construction and sequencing from pooled versus separate samples was among the first issues debated by the team at the beginning of the project. The numbers of time points for each stress treatment and the desire to start with detailed time-course experiments for each stress and each species was also debated at the beginning. For practical concerns, a more simple strategy was selected, despite our own desire to have a more robust data set. Hopefully the following explanation articulates why.

Our grant proposal, as originally written, included roughly equal parts of genomic resource development and hypothesis driven scientific enquiry. **Our project was ultimately approved as a resource development grant only** at half of the budget proposed, and the research studies were removed at the request of NSF Plant Genome Research Program. In our pre-award negotiations with NSF, the team decided that it was most important to develop EST databases and mapping populations for as many species as possible as highest priorities, and to keep the project team together. We also decided to retain the original grant title as that aligned with our long term interests, such as obtaining a follow-up, "part 2" grant to address our scientific interests. But **we admit it was a key error to not revise the title of the grant**, which continues to imply that a major emphasis of the project is comparative genomics of environmental stress

response science, rather than just resource development. We attempted to retain the environmental stress component of our original vision by acquiring transcriptome sequences from stress treated seedlings. However that has meant pooling samples rather than not benefitting from appropriate biological replications, mainly from budget constraints. Another practical constraint was that the stress treatments could only be conducted with seedlings, and only one or two leaves could be sampled at each time point from the small one to two year old seedlings for these species. RNA preparations need to be of sufficient quantity and at high purity for Next Gen sequencing. We found it was not possible to get sufficient amounts and purity of RNA using only one or two leaves from each seedling. This may in part be because the stress conditions themselves caused an overall reduction in transcription and/or degradation of existing transcripts. We were able to obtain much better mRNA preparations by pooling leaves from among the individual seedlings at each time point prior to RNA preparation.

We continue to hope that the data from each stress from multiple species will provide interesting preliminary data, and a resource, upon which future hypothesis-testing research projects can be based, even though we recognize the limitations in not being able to compare among individuals within species, we are limited in what we can deduce from the between species comparisons. The SAB suggested that we repeat one set of stress experiments, preparing and sequencing individual, non-pooled libraries for comparison to the data from pooled samples. This way, the amount of variation among plants could be assessed and the amount of information lost in the pooled sample approach could be determined. We have chosen to conduct such a study ozone treatments, because of the interesting results we are obtaining in comparing ozone treatments among species. **We will conduct a new set of ozone treatments with northern red oak half-sib seedlings in year 4 and try to prepare and sequence individual RNA libraries for treated leaves from individual plants for at least two time points and two ozone levels.** The other time points and treatment levels may be from pooled libraries. We hope this will allow us to do the suggested comparisons between pooled and individual RNASeq data sets. In addition, it will provide Illumina data for red oak ozone treatments which will aid in our between species comparisons, as the data obtained in year one from red oak was from 454 sequencing and only half the depth of Illumina data obtained for the other species.

Another suggestion of the SAB is to repeat our original stress treatments with red oak and black walnut full-sib seedlings that failed due to handling and shipping issues, and we agree these should be redone. The full-sib oak and walnut seedlings that we used are now out-planted in in field. However, we can use half sib seedlings of the appropriate age for both species even though somewhat less comparable to the full sib seedlings used with the other species. Unfortunately, after detailed study of our budget, **our funds for sequencing in the final year are sufficient only to complete the above work and to develop EST datasets for the final two remaining species** (sugar maple and sweetgum), in the manner completed for previous species. We do not have the funds remaining in our budget to also do RNASeq of additional stress treated red oak and black walnut seedlings. One reason for the budget shortfall is **that in years**

2 and 3, we conducted deep transcriptome sequencing of multiple tissue libraries from parental trees of our mapping populations for both walnut and red oak, to replace the EST data lost from the failed stress treatments in year 1. The parental tissues yielded 45.1 Gb of new EST data for *Quercus rubra* and 47.9 Gb for Juglans nigra which permitted us to move ahead with gene annotation and SSR and SNP marker development for these key species. This deep EST data is now a key resource for the SSR and GBS mapping work. However, although we cannot afford additional RNA sequencing, **we will repeat all of the stress treatments with half sib seedlings of red oak and black walnut in year 4 with and preserve the tissues from each plant separately** at -80C, for individual RNASeq libraries to be sequenced in subsequent projects. Finally, **we will insure that treated and control plants are adequately segregated** in all of our final year's experiments, as suggested by the SAB.

Our sincere thanks again to the SAB for the valuable advice and feedback on our project. We are looking forward to seeing you again at our final annual meeting at Michigan Tech University in August of 2014. Nick Wheeler will be in touch with the dates and arrangements.